THE ANNALS *of* MATHEMATICAL STATISTICS

(FOUNDED BY H. C. CARVER)

THE OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

Contents

	PAGE
On The Theory of Systematic Sampling, II. William G. MADOW	333
Problems in Plane Sampling. M. H. QUENOUILLE	355
Representation of Probability Distributions by Charlier Series. R. P. BOAS, JR.	376
Heuristic Approach to the Kolmogorov-Smirnov Theorems. J. L. DOOB	393
Pearsonian Correlation Coefficients Associated with Least Squares Theory. PAUL S. DWYER	404
Inversion Formulas in Normal Variable Mapping. JOHN RIORDAN	417
On the Determination of Optimum Probabilities in Sampling. MORRIS H. HANSEN AND WILLIAM N. HURWITZ	426
A Solution to the Problem of Optimum Classification. P. G. HOEL AND R. P. PETERSON	433
Notes:	
A Generalization of Wald's Fundamental Identity. GUNNAR BLOM	439
Spread of Minima of Large Samples. Brockway McMILLAN	444
On the Convergence of the Classical Iterative Method of Solving Linear Simultaneous Equations. EDGAR REICH	448
Some Recurrence Formulae in the Incomplete Beta Function Ratio. T. A. BANCROFT	451
On a Theorem by Wald and Wolfowitz. GOTTFRIED E. NOETHER	455
On Sums of Symmetrically Truncated Normal Random Variables. Z. W. BIRNBAUM AND F. C. ANDREWS	458
A Certain Cumulative Probability Function. SISTER MARY AGNES HATKE	461
Abstracts of Papers	464
News and Notices	470
Report on the Berkeley Meeting of the Institute	475

Vol. XX, No. 3 — September, 1949

THE ANNALS OF MATHEMATICAL STATISTICS

The ANNALS OF MATHEMATICAL STATISTICS is published quarterly by the Institute of Mathematical Statistics, Mt. Royal & Guilford Aves., Baltimore 2, Md. Subscriptions, renewals, orders for back numbers and other business communications should be sent to the ANNALS OF MATHEMATICAL STATISTICS, Mt. Royal & Guilford Aves., Baltimore 2, Md., or to the Secretary of the Institute of Mathematical Statistics, C. H. Fischer, Business Administration Building, University of Michigan, Ann Arbor, Mich.

Changes in mailing address which are to become effective for a given issue should be reported to the Secretary on or before the 15th of the month preceding the month of that issue. The months of issue are March, June, September and December.

Manuscripts for publication in the ANNALS OF MATHEMATICAL STATISTICS should be sent to S. S. Wilks, Fine Hall, Princeton, New Jersey. Manuscripts should be typewritten double-spaced with wide margins, and the original copy should be submitted. Footnotes should be reduced to a minimum and whenever possible replaced by a bibliography at the end of the paper; formulae in footnotes should be avoided. Figures, charts, and diagrams should be drawn on plain white paper or tracing cloth in black India ink twice the size they are to be printed. Authors are requested to keep in mind typographical difficulties of complicated mathematical formulae.

Authors will ordinarily receive only galley proofs. Fifty reprints without covers will be furnished free. Additional reprints and covers furnished at cost.

The subscription price for the ANNALS is $8.00 inside the Western Hemisphere and $5.00 elsewhere. Single copies $3.00. Back numbers are available at $8.00 per volume or $3.00 per single issue.

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD., U. S. A.

Entered as second-class matter at the Post Office at Baltimore, Maryland, under the act of March 3, 1879.

ON THE THEORY OF SYSTEMATIC SAMPLING, II

By William G. Madow[1]

Institute of Statistics, University of North Carolina

1. Summary and introduction. In an earlier paper,[2] [1] an approach to the problem of systematic sampling was formulated, and the associated variance obtained. Several forms of the population were assumed. The efficiency of the systematic design as compared with the random and stratified random design was evaluated for these forms. It was remarked that as the size of sample increased the variance of a systematic design might also increase, contrary to the behavior of variances in the random sampling design. This possibility was verified in [2].

One approach to the study of systematic designs, given by Cochran [3] removed this difficulty to some extent by changing the problem to one of the expected variance, and supposing the elements of the population to be random variables. He showed that if the correlogram of these random variables is concave upwards, then the expected variance of the systematic design would be less, and often considerably less, than the variance of a stratified design.

In the present paper the results of the earlier papers are extended to the systematic sampling of clusters of equal and unequal sizes. Some comments on systematic sampling in two dimensions are included.

In section 2 we derive two theorems that have considerable applications in many parts of sampling. Although it has been common for people working in sampling theory to tell each other that these theorems ought to be true, yet no reference seems to exist.

In section 3 we develop the implications of a remark [1, p. 13] that in designing sample surveys we should try to induce negative correlation between strata. In Theorem 3 we obtain sufficient conditions for the correlation to be negative. The lemma and Theorem 4 given in Section 4 enable us to extend the uses of Theorem 3 in practice. As an application of these results, we show that if a population has a concave upwards correlogram, and if strata are defined in an optimum fashion for the selection of one element at random from each stratum, then we can define a systematic type design that will be more efficient than independent random selection from each stratum.

In sections 5 and 6 we obtain various results in the systematic sampling of clusters largely as applications of the more general theorems of the earlier sections. In general the results are of a nature similar to those of [1] and [3] in that the formulae show the conditions under which systematic sampling may be expected to be more efficient than random or stratified random sampling. We have not, however, applied these formulae to specified types of populations.

[1] Submitted for publication, November, 1948. Parts of this paper were prepared while the author was Visiting Professor of Statistics at the University of São Paulo, Brazil.

[2] References to the articles and book cited are given by Roman numerals.

From [1, 2 and 3] it is already apparent that this work will be useful and such studies should be more valuable when made in connection with important types of surveys or data than when made as illustrations in a general paper.

2. Random events and conditional expectations. Almost invariably, samples are selected in several stages. For example, to select a sample of households from a city one frequently used method is the following two stage sampling plan:

a. A map of the city showing the location of each block is obtained and brought up-to-date.
b. Using this map, a sample of the blocks of the city is selected (this is stage1).
c. From the households on the blocks selected in stage 1, a subsample of households is selected (this is stage 2.).

In this section, we give a general approach for evaluating the means and variances associated with multi-stage sampling. This approach has the advantage of at once yielding the contributions to the variance arising from each stage. Furthermore, the theorems presented are useful in calculating variances even when our interest is not in multi-stage sampling. The theorems are presented in general terms because of their wide application in sampling.

We shall say that the result of performing an operation is a random event A^* if the result can assume m possible states $A_1, \cdots, A_m$ with probabilities $p_1, \cdots, p_m$, where

$$P\{A^* = A_i\} = p_i, \qquad \sum_{i=1}^{m} p_i = 1,$$

and $P\{A^* = A_i\}$ is read "the probability that the random event A^* assumes the state A_i."

One illustration of an operation is the operation of selecting a sample of blocks. If there are N blocks in the city of which we select n in such a way that each set of n of the N blocks is a possible sample, then there are C_n^N possible samples. In this case $m = C_n^N$ and the C_n^N possible samples are the m states of A^* "the result of selecting the sample of blocks." Furthermore, if each of the possible samples of blocks is equally likely to be selected, then

$$P\{A^* = A_i\} = \frac{1}{C_n^N} = \frac{1}{m}.$$

The random event A^* may also be the taking on by a random variable of one of its possible values. If z^* is a random variable having possible values $z_1, \cdots, z_m$ with probabilities $p_1, \cdots, p_m$ then we can define the states of A^* to be A_i where A_i is "$z^* = z_i$."

Thus the notion of a random event includes the two types of randomness that are met in selecting samples.

Let x' be a random variable. Then, by the conditional expectation of x' subject to the random event A^* is meant the random variable $E^*(x' \mid A)$ whose possible values are $E(x' \mid A_i)$, $i = 1, \cdots, m$ and whose probabilities are p_i, that is

$$P\{E^*(x' \mid A) = E(x' \mid A_i)\} = p_i = P\{A^* = A_i\},$$

where

$$E(x' \mid A_i) = \sum_{j=1}^{N_i} x_{ij} p_j(A_i), \tag{2.1}$$

x_{ij} is the jth of the N_i possible values of x' when A_i occurs, and

$$p_j(A_i) = P\{x' = x_{ij} \mid A_i\}$$

is "the probability that $x' = x_{ij}$ given that A_i occurs." It should be noted that if

$$p_{ij} = P\{x' = x_{ij}\},$$

then

$$p_{ij} = P\{x' = x_{ij}, A^* = A_i\}$$

since the fact that $x' = x_{ij}$ implies the occurrence of A_i. Then

$$p_i \cdot p_j(A_i) = p_{ij}. \tag{2.2}$$

We state Theorems 1 and 2 without proof since their proofs are immediate.

THEOREM 1. *The expected value of the random variable $E^*(x' \mid A)$ is $E\,x'$,* i.e.

$$E\{E^*(x' \mid A)\} = Ex'.$$

By $\sigma^*_{x'y' \mid A}$ we shall mean the random variable whose possible values are $\sigma_{x'y' \mid A_i}$, $i = 1, \cdots, m$ where

$$\sigma_{x'y' \mid A_i} = E\{[x' - E(x' \mid A_i)]\,[y' - E(y' \mid A_i)] \mid A_i\}$$

and

$$P\{\sigma_{x'y' \mid A} = \sigma_{x'y' \mid A'}\} = p_i = P\{A^* = A_i\},$$

i.e.

$$\sigma^*_{x'y' \mid A} = E^*\{[x' - E^*(x' \mid A)]\,[y' - E^*(y' \mid A)] \mid A\}.$$

Furthermore, the symbol $\sigma_{E^*(x' \mid A)\,E^*(y' \mid A)}$ will stand for "the covariance of the two random variables $E^*(x' \mid A)$ and $E^*(y' \mid A)$." The corresponding definitions of variance are obtained by replacing y' by x' above.

THEOREM 2. *If x' and y' are random variables, then*

$$\sigma_{x'y'} = E\sigma^*_{x'y' \mid A} + \sigma_{E^*(x' \mid A)\;E^*(y' \mid A)}$$

and

$$\sigma^2_{x'} = E\sigma^2_{x' \mid A} + \sigma^2_{E^*(x' \mid A)}.$$

We note that, since the p_{ij}, p_i and $p_j(A_i)$ are not specified, Theorems 1 and 2 are valid for any two-stage plan. The generalizations of Theorems 1 and 2 to multi-stage plans are obvious, but in practice it often turns out to be simpler to apply the theorems several times.

It would be easy to give applications of Theorems 1 and 2 but these are not essential for our purposes in this paper. As remarked in the introduction, these two theorems have long been part of what we may call the folklore of sampling.

3. Stratified sampling and negative correlation, with an application to systematic sampling. In discussing plans for sampling from a stratified population it is customary to suppose that if x' is an estimate and $x' = x'_1 + \cdots + x'_L$ where x'_i is the contribution to x' arising from the jth of the L strata, then the sampling is to be so done that the random variables x'_i and x'_j, $j \neq i$, are independent.

In [1, p. 13] it was noted that if a population were stratified, and if the elements were so selected that the contributions from different strata were negatively correlated, it would follow that the variance of the estimate would be less than if the contributions were independent but had the same covariances within strata. This was, of course, an immediate conclusion from the fact that

$$\sigma^2_{x'} = \sum_{i,j=1}^{L} \sigma_{x'_i x'_j}$$

and, hence, if

$$C = \sum_{i \neq j} \sigma_{x'_i x'_j} < 0 \tag{3.1}$$

then $\sigma^2_{x'}$ is less than it would be if $C = 0$. If $C < 0$ we shall say that the sample design has "negative correlation."

It is obvious that any population may be taken to be itself a sample, a sample from the possible populations that might have been produced by the forces that determined the existing population. Inasmuch as sampling designs are often chosen on the basis of a knowledge of the dominating forces and some past experience, it is realistic to consider not only the expected values and variances for a specific population but also their expected values over all possible populations determined by the same forces. Cochran [3] has given one illustration of the usefulness of considering the expected variance of a sample design. He considered the elements $x_1, \cdots, x_n$ of the population themselves to be random variables and supposed that $E_i x_i = \mu$ and $E(x_i - \mu)^2 = \sigma^2$. For his purposes it was also convenient to suppose that if $u > 0$ then $E(x_i - \mu)(x_{i+u} - \mu) = \rho_u \sigma^2$. It was then possible for him to make realistic hypotheses concerning the correlogram, i.e. the ρ_u considered as a function of u, that would not have been reasonable in dealing with a specific population. He thus obtained general conclusions concerning the expected efficiency of systematic sampling designs as compared with random and stratified random designs.

In this paper we shall consider not only the expected values and variances for the given finite population but also the expected values of these expected values and variances under the assumption that the elements of the population are themselves random variables. We shall use $\mathcal{E}$ to denote the expected value

considering the elements of the population to be random variables and as before use E for expected values based on the specified finite population.

Then

$$\mathcal{E}\sigma_{x'}^2 = \sum_{i,j=1}^{L} \mathcal{E}\sigma_{x_i' x_j'},$$

and if $\mathcal{E}C < 0$ we shall say that the design has 'expected negative correlation.'

We now propose to obtain the beginnings of an approach to sample design when it is possible to introduce or take advantage of negative correlation or expected negative correlation through the sample design.

To simplify, we shall begin by considering two strata and shall suppose that the possible values of x' are $x_1, \cdots, x_n$ while the possible values of y' are $y_1, \cdots, y_n$. Furthermore, we shall suppose the sampling to be so done that

$$P\{x' = x_i\} = P\{y' = y_i\} = P\{x' = x_i, y' = y_i\} = p_i > 0,$$

so that $\sum_{i=1}^{n} p_i = 1$ and $P\{x' = x_i, y' = y_j\} = 0$ if $i \neq j$.

Under the above assumptions, it follows that

$$\sigma_{x'y'} = \sum_{i=1}^{n} p_i x_i y_i - \sum_{i,j=1}^{n} p_i p_j x_i y_j . \tag{3.2}$$

The symbol $\varphi_{ij} \gtrsim 0$ means that $\varphi_{ij} \geq 0$ for all i and j and $\varphi_{ij} > 0$ for at least one pair i, j. We shall say that if $(x_i - x_j)(y_i - y_j) \gtrsim 0$ then the sets (x) and (y), where (x) stands for $x_1, \cdots, x_n$ and (y) for $y_1, \cdots, y_n$ are similarly ordered and if $(x_i - x_j)(y_i - y_j) \lesssim 0$ then these sets are oppositely ordered. Then it is easy to prove, [4, p. 43] directly that if the values are oppositely ordered, then $\sigma_{x'y'} < 0$ and if they are similarly ordered then $\sigma_{x'y'} > 0$.

A somewhat more general result is the following:

THEOREM 3. *Let $n \leq k$, let*

$$b = \sum_{i=1}^{n} \sum_{j=1}^{k} a_{ij} w_i z_j$$

be a real bilinear form, and let

$$t = \sum_{i=j}^{n} a_{ii} w_i$$

be a real linear form, where $w_i > 0$, $z_i > 0$ and $\sum_{i=1}^{n} w_i = \sum_{i=1}^{k} z_i = 1$.

Then a sufficient condition that $b > t$ is

$$a_{ij} \gtrsim a_{ij} . \tag{3.3}$$

If $k = n$ and $w_i = z_i$ then $b > t$ if

$$a_{ij} + a_{ji} \gtrsim a_{ii} + a_{jj} . \tag{3.4}$$

PROOF. Since

$$b - t = \sum_{i=1}^{n} a_{ii}(w_i z_i - w_i) + \sum_{i \neq j} a_{ij} w_i z_j ,$$

and since

$$1 - z_i = \sum_{\substack{j=1 \\ j \neq i}}^{k} z_j ,$$

it follows that

$$b - t \sum_{j \neq i} (a_{ij} - a_{ii}) w_i z_j .$$

Hence, $b > t$ if (3.3) holds. Also, if $k = n$ and $w_i = z_i$ then $b > t$ if (3.4) holds. Some obvious generalizations of Theorem 3 have been omitted since we do not need them.

To obtain the result that $\sigma_{x'y'} < 0$ if the sets (x) and (y) are oppositely ordered, we make the identifications $a_{ij} = x_i y_j$ and $z_i = w_i = p_i$. Then (3.4) holds and substituting we have

$$a_{ii} + a_{jj} - a_{ij} - a_{ji} = (x_i - x_j)(y_i - y_j) \tag{3.5}$$

so that if the values are oppositely ordered, $\sigma_{x'y'} < 0$, and hence the two strata have negative correlation.

To consider expected negative correlation we note that

$$\mathcal{E}\sigma_{x'y'} = \sum_{i=1}^{n} p_i \sigma_{ii} + \sum_{i,j=1}^{n} p_i p_j \sigma_{ij} \tag{3.6}$$

where we suppose that $\mathcal{E}x_i = \mu$, $\mathcal{E}y_i = \nu$ and

$$\mathcal{E}(x_i - \mu)(y_j - \nu) = \sigma_{ij}$$

so that in this case σ_{ii} is a covariance, not a variance.

If we put $a_{ij} = \sigma_{ij}$ and $z_i = w_i = p_i$, then (3.4) holds and we obtain, as sufficient for $\mathcal{E}\sigma_{x'y'}$ to be negative, that

$$\sigma_{ij} + \sigma_{ji} \geq \sigma_{ii} + \sigma_{jj} \tag{3.7}$$

or, if we define ρ_{ij} by the equation,

$$\sigma_x \sigma_y \rho_{ij} = \sigma_{ij} ,$$

where $\sigma_x^2 = \mathcal{E}(x_i - \mu)^2$ and $\sigma_y^2 = \mathcal{E}(y_i - \nu)^2$, we have

$$\rho_{ij} + \rho_{ji} \geq \rho_{ii} + \rho_{jj} \tag{3.8}$$

as a sufficient condition for $\mathcal{E}\sigma_{x'y'} < 0$.

Let us consider the systematic sampling of single elements. In systematic sampling, we assume a population of kn ordered elements $x_1, x_2, \cdots, x_k, x_{1+k}, \cdots, x_{2k}, \cdots, x_{1+(n-1)k}, \cdots, x_{nk}$ of which we wish to estimate the arith-

metic mean $\bar{x}$. As our estimate we use

$$\bar{x}' = (x_1' + \cdots + x_m')/n$$

where x_1' is selected at random from $x_1, \cdots, x_k$ and if $x_1' = x_j$ then $x_i' = x_{j+(i-1)k}$, $i = 2, \cdots, m$. Thus, $\bar{x}'$ may be interpreted as an estimate based on a stratified population, the ith stratum consisting of

$$x_{1+(i-1)k}, \cdots, x_{k+(i-1)k}$$

and

$$P\{x_i' = x_{\alpha+(i-1)k}\} = P\{x_i' = x_{\alpha+(i-1)k}, x_j' = x_{\alpha+(j-1)k}\} = 1/k$$

while

$$P\{x_i' = x_{\alpha+(i-1)k}, x_j' = x_{\beta+(j-1)k}\} = 0, \text{ if } \alpha \neq \beta.$$

Then

$$\sigma_{x_i'x_j'} = \left(\frac{1}{k}\right) \sum_{\alpha=1}^{k} x_{\alpha+(i-1)k} \cdot x_{\alpha+(j-1)k} - \bar{x}_i \bar{x}_j$$

where

$$\bar{x}_i = \left(\frac{1}{k}\right) \sum_{\alpha=1}^{k} x_{\alpha+(i-1)k}.$$

Hence, any two strata that are oppositely ordered will yield a negative contribution to the variance. However, since it is not possible for all strata to be negatively ordered, we do not thus obtain a useful result and must return to the consideration of C or $\sigma_{\bar{x}'}^2$ itself as was done in [1]. If, however, we make Cochran's assumptions, and consider $\mathcal{E}\sigma_{x'y'}$, it follows that for the ith and jth strata

$$\rho_{\alpha\beta} = \rho_{(j-i)k+\beta-\alpha},$$

and (3.8) becomes

$$\rho_{(j-i)k+(\beta-\alpha)} + \rho_{(j-i)k+(\alpha-\beta)} \geq 2\rho_{(j-i)k}, \tag{3.9}$$

i.e. the correlation function ρ_u must be concave upwards, which Cochran showed by other means. By considering $\mathcal{E}C$ it is possible to show that a sort of average concavity is all that is required of the correlogram for systematic sampling to have a smaller variance than stratified random sampling.

4. Conditions for negative correlation when the strata are of unequal sizes with an application to systematic sampling. Often, as in the systematic selection of clusters with probability proportionate to size (discussed in Section 5) the simplified situation dealt with in Theorem 3 does not directly apply. However, Theorem 3 may be used to advantage by the following device.

Let us suppose the possible values of x' to be $x_1, \cdots, x_n$ and those of y_0' to be $y_1^0, \cdots, y_k^0$, $k > n$ and let

$$P\{y' = y_\beta^0 \mid x' = x_\alpha\} = p_{\beta|\alpha}$$

so that if we define

$$y_\alpha = \sum_{\beta=1}^{k} y_\beta^0 p_{\beta|\alpha}, \tag{4.1}$$

then

$$y_\alpha = E(y_0' \mid x' = x_\alpha).$$

If we define y' to be a random variable having possible values $y_1, \cdots, y_n$ with probabilities $p_1, \cdots, p_n$ where

$$p_\alpha = P\{x' = x_\alpha\}$$

it follows that

$$y' = E^*(y_0' \mid x')$$

and

$$\sigma_{x'y_0'} = \sigma_{x'y'}.$$

Clearly, Theorem 3 is valid for the random variables x' and y'.

Consequently, we need only determine what restrictions the conditional probabilities, $p_{\beta|\alpha}$, and the values, y_α^0, need satisfy for the sets $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$ to be oppositely ordered or for (3.7) to hold.

Substituting for y_i and y_j in (3.5) we see that if

$$(x_\alpha - x_\gamma) \sum_{\beta=1}^{k} y_\beta^0 (p_{\beta|\alpha} - p_{\beta|\gamma}) \leqq 0 \tag{4.2}$$

then $\sigma_{x'y'} = \sigma_{x'y_0'} < 0$.

Let

$$\sigma_{\alpha\gamma}^0 = \mathfrak{E}(x_\alpha - \mu)\,(y_\gamma^0 - \nu).$$

Then substituting in (3.7) we see that if

$$\sum_{\beta=1}^{k} (p_{\beta|\alpha} - p_{\beta|\gamma})(\sigma_{\alpha\beta}^0 - \sigma_{\gamma\beta}^0) \leqq 0 \tag{4.3}$$

or if

$$\sum_{\beta=1}^{k} (p_{\beta|\alpha} - p_{\beta|\gamma})(\rho_{\alpha\beta}^0 - \rho_{\gamma\beta}^0) \leqq 0 \tag{4.4}$$

then

$$\mathfrak{E}\sigma_{x'y_0'} < 0.$$

In order to use (4.2) and (4.3) the following well-known lemma is often useful.

LEMMA. *If* $\xi_1 \leq \xi_2 \leq \cdots \leq \xi_k \leq 0$ *and the quantities* $\epsilon_1, \cdots, \epsilon_k$ *are such that*

$$\sum_{\beta=1}^{s} \epsilon_\beta \geq 0$$

then

$$\sum_{\beta=1}^{s} \epsilon_\beta \xi_\beta \leq 0, \qquad s = 1, \cdots, k.$$

Let us use this lemma to obtain another theorem that will be helpful in showing negative or expected negative correlation between strata.

THEOREM 4. *Let b be a bilinear form*

$$b = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} w_i z_j$$

such that $\sum_{i=1}^{s} w_i \geq 0$, $\sum_{j=1}^{s'} z_j \geq 0$, $s = 1, \cdots, n-1$, $s' = 1, \cdots, m-1$, *and*

$$\sum_{i=1}^{n} w_i = \sum_{j=1}^{m} z_j = 0. \tag{4.5}$$

Let

$$\delta_{ij} = a_{ij} - a_{i+1,j} - a_{i,j+1} + a_{i+1,j+1}.$$

Then a sufficient condition that $b \leq 0$ is $\delta_{ij} \leq 0$.

PROOF. Upon substituting for w_n and z_m in b from (4.5) we see that

$$b = \sum_{i=1}^{n-1} \sum_{j=1}^{m-1} \delta'_{ij} w_i z_j$$

where

$$\delta'_{ij} = a_{ij} - a_{im} - a_{nj} + a_{nm}$$

or, if we define,

$$\xi_j = \sum_{i=1}^{n-1} \delta'_{ij} w_i$$

then

$$b = \sum_{j=1}^{m-1} \xi_j z_j.$$

According to the lemma, it then follows that a sufficient condition that $b \leq 0$ is that

$$\xi_1 \leq \xi_2 \leq \cdots \leq \xi_{m-1} \leq 0.$$

Also, a sufficient condition that

$$\xi_j - \xi_{j+1} \leq 0,$$

is

$$\delta'_{ij} - \delta'_{i,j+1} \leq \delta'_{i+1,j} - \delta'_{i+1,j+1}$$

Then to complete the proof it is only necessary to verify that

$$\delta_{ij} = \delta'_{ij} - \delta'_{i,j+1} - \delta'_{i+1,j} + \delta'_{i+1,j+1} .$$

In the preceding pages we have given an identification of systematic with stratified sampling where, instead of the selection being made independently within strata, the choice of an element from one stratum determines the choice from the other strata. In this identification, however, it was assumed that the strata contained the same number of elements. Let us now extend this method of selecting samples to the case where the strata have different numbers of elements. In so doing we shall illustrate the use of the above lemma and theorem 4.

Suppose now that the population consists of N elements $x_1 , \cdots , x_N$ classified into n strata, the ith of which contains the N_i elements

$$x_{N_1+\cdots+N_{i-1}+1} , \cdots , x_{N_1+\cdots+N_i} .$$

We shall denote these elements by $x_{i1} , \cdots , x_{iN_i}$.

We shall select one element from each of these n strata. The element selected from the ith stratum is written x'_i . As the estimate of $\bar{x}$, the arithmetic mean of the population, we use

$$\bar{x}' = \sum_{i=1}^{n} \frac{N_i}{N} x'_i$$

and it is well known that if the selection is made independently at random from each stratum, then

$$\sigma^2_{\bar{x}'} = \sum_{i=1}^{n} \left(\frac{N_i}{N}\right)^2 \sigma_i^2$$

where σ_i^2 is the variance of x'_1 , i.e. the variance of the ith stratum.

Let us now consider an alternative to the usual method. We can suppose that $N_1 > 1$ without any loss of generality. (The methods are the same for any stratum having $N_i = 1$ and will also yield the same result for any population such that either all the $N_i = 1$ or all but one of the $N_i = 1$. Differences occur if at least two of the N_i differ from 1.)

We first choose an element at random from the first stratum. Suppose that $x'_1 = x_\alpha$. Then to choose an element from the second stratum, assuming that $N_2 > 1$, we proceed as follows: Multiply N_2 by any positive integer t_2 such that $N_2 t_2 / N_1$ is an integer, say, k_2 . Assign to each element of the second stratum the measure of size t_2 , and form the two sets of cumulative totals $t_2 , 2t_2 , \cdots , N_2 t_2$ and $k_2 , 2k_2 , \cdots , N_1 k_2$. Then with the measures of size t_2 assigned to each element of stratum 2, and the measure of size k_2 assigned to each element of stratum 1, it follows that strata 1 and 2 have the same total size.

As an example of the arithmetic given below consider the following simple case. Suppose that $N_1 = 3$ and $N_2 = 4$. Then if we take for t_2 the value 6, it follows that $k_2 = 8$. We choose one of the integers 1, 2, 3 with equal probability. If the

integer 1 is obtained, we have selected the first element of the first stratum and choose an integer between 1 and 8 with equal probability. If the selected integer is between 1 and 6, the first element of the second stratum is selected. If it is 7 or 8 the second element of the second stratum is selected. Similarly if the second element of the first stratum is selected, then we select an integer between 9 and 16 with equal probability. If that integer has value 9, $\cdots$, 12 the second element of the second stratum is selected; if it has value 13, $\cdots$, 16 the third element is selected.

The general formulation of the selection procedure for the second stratum is:

Suppose that β_0 is the smallest integer such that $(\alpha - 1)k_2 + 1 \leq \beta_0 l_2$ and that β_1 is such that $(\beta_1 - 1)l_2 < \alpha k_2 \leq \beta_1 l_2$. Choosen an integer at random from $1, \cdots, k_2$ and call that integer β. Then, if

$$(\alpha - 1)k_2 < (\alpha - 1)k_2 + \beta \leq \beta_0 l_2$$

the β_0th element is selected from stratum 2; if

$$\beta_0 l_2 < (\alpha - 1)k_2 + \beta \leq (\beta_0 + 1)l_2$$

the $(\beta_0 + 1)$th element is selected; $\cdots$; and if

$$(\beta_1 - 1)l_2 < (\alpha - 1)k_2 + \beta \leq \alpha k_2$$

the β_1th element is selected from stratum 2.

It is easy to verify that when the sample is so selected, each element of stratum 2 has equal probability of being selected. Hence, if we apply this procedure to each stratum we have

$$\sigma_{\bar{x}'}^2 = \sum_{i=1}^{n} \left(\frac{N_i}{N}\right)^2 \sigma_i^2 + \sum_{i \neq j} \frac{N_i N_j}{N^2} \sigma_{x_i' x_j'}.$$

Let us evaluate $\sigma_{x_i' x_j'}$ for this type of selection. Now

$$\sigma_{x_i' x_j'} = E\,(x_i' - \bar{x}_i)(x_j' - \bar{x}_j)$$

where $\bar{x}_i$ is the arithmetic mean of the elements of the ith stratum. From Theorem 2, we then have

$$\sigma_{x_i' x_j'} = \frac{1}{N_1} \sum_{\alpha=1}^{N_1} E[x_i' - E(x_i' \mid x_{1\alpha})][x_j' - E(x_j' \mid x_{1\alpha}]$$

$$+ \frac{1}{N_1} \sum_{\alpha=1}^{N_1} [E(x_i' \mid x_{1\alpha}) - \bar{x}_i][E(x_j' \mid x_{1\alpha}) - \bar{x}_j].$$

It is easy to see that the method of selection used above implies that the first term of $\sigma_{x_i' x_j'}$ vanishes. Furthermore, $\bar{x}_i$ is the arithmetic mean of the conditional expectations so that we have reduced the problem to one of determining whether the conditional expectations satisfy the conditions for negative correlation or expected negative correlation.

If we denote $E\,(x_i' \mid x_{1\alpha})$ by $y_{i\alpha}$, then we need to see whether the sets $y_{i1} \cdots$,

y_{iN_1} and $y_{j1}, \cdots, y_{jN_1}$ are oppositely ordered. Now

$$(y_{i\alpha} - y_{i\beta})(y_{j\alpha} - y_{j\beta}) = \sum_{g=1}^{N_i} \sum_{h=1}^{N_j} x_{ig} x_{jh} \epsilon_{ig\alpha\beta} \epsilon_{jh\alpha\beta}$$

where

$$\epsilon_{ig\alpha\beta} = P\{x_i' = x_{ig} \mid x_{1\alpha}\} - P\{x_i' = x_{ig} \mid x_{1\beta}\}.$$

If $\alpha < \beta$ then, according to the method of selection,

$$\sum_{g=1}^{s} \epsilon_{ig\alpha\beta} \geq 0, \qquad s = 1, \cdots, N_i - 1$$

while

$$\sum_{g=1}^{N_i} \epsilon_{ig\alpha\beta} = 0.$$

In Theorem 4, we then make the identifications $n = N_i$, $m = N_j$,

$$w_g = \epsilon_{ig\alpha\beta}, \qquad z_h = \epsilon_{jh\alpha\beta} \quad \text{and} \quad a_{gh} = x_{ig} x_{jh}.$$

Then

$$\delta_{gh} = (x_{ig} - x_{i,g+1})(x_{jh} - x_{j,h+1})$$

and hence to have negative correlation between the strata, it is sufficient that the sets $x_{i1}, \cdots, x_{iN_i}$ and $x_{j1}, \cdots, x_{jN_j}$ have the type of negative ordering represented by $\delta_{gh} \leq 0$. Similarly, if

$$\sigma_{gh} = \mathfrak{E}(x_{ig} - \mu_i)(x_{jh} - \mu_j), \qquad \mu_i = \mathfrak{E}x_{ig},$$

then, for expected negative correlation, it is sufficient that

$$\sigma_{gh} - \sigma_{g,h+1} - \sigma_{g+1,h} + \sigma_{g+1,h+1} \leq 0.$$

Of course, these conditions will be satisfied if a concave upwards correlogram exists. Hence, if a population consists of N random variables $x_1, \cdots, x_N$ having a concave upwards correlogram, then, no matter into what strata these elements are classified, provided that the order of occurrence of the elements remains unaltered, the systematic selection of the elements in the sample can be so planned as to yield an estimate having smaller variance than the stratified random selection of the elements in the sample even if optimum allocation is used. If more than one element is being selected from a stratum under optimum allocation, then the systematic selection of the same number of elements will suffice. If not only optimum allocation but also optimum definitions of strata are being used so that but one element is selected from each stratum, then systematic selection according to the scheme described in the section will produce a variance not larger than the variance of stratified random sampling. It should be noted, however, that this does not imply that a 'hammer and tongs' use of systematic sampling ignoring the strata will produce a smaller variance. There is work to be done on what is required for the latter to occur.

It may be noted that the procedure of this example provides an answer to the systematic selection of elements from a population whose size is not a multiple of the size of sample.

5. The systematic sampling of clusters with probability proportionate to a measure of size. It is known [5] that sampling clusters with probability proportionate to a measure of size often yields considerable reductions in the variance of the estimates. However, the theory of the systematic selection of several clusters with probability proportionate to a measure of size has not been worked out, and it is the purpose of this section to make some contributions to that theory.

The most frequently used method of sampling clusters with probability proportionate to size is equivalent to the following: Suppose that the clusters are denoted by $C_1, \cdots, C_M$ and that to the hth of these M clusters is assigned a measure of size P_h. Form the successive totals $P_1, P_1 + P_2, P_1 + P_2 + P_3, \cdots, P_1 + \cdots + P_M$. If we wish to select m of these clusters, we calculate $\overline{P}_m = (P_1 + \cdots + P_M)/m$. Then, assuming that $P_j \leq \overline{P}_m$, $j = 1, \cdots, M$, we select an integer with equal probability from $1, \cdots, \overline{P}_m$. Calling that integer P', we calculate the m numbers $P', P' + \overline{P}_m, P' + 2\overline{P}_m, \cdots, P' + (m-1)\overline{P}_m$. If

$$P_1 + \cdots + P_{h-1} + 1 \leq P' + (i-1)\overline{P}_m \leq P_1 + \cdots + P_h \tag{5.1}$$

for any integer i, $i = 1, \cdots, m$, then the cluster C_h is selected for the sample. Any cluster for which $P_h > \overline{P}_m$ is automatically included in the sample, and if there are, say, α such clusters, then we calculate $\overline{P}_{m-\alpha}$ for the $M - \alpha$ clusters remaining after including these α in the sample, and proceed as above.

In deriving the variance of the estimate we shall use, we interpret that estimate as a stratified sampling estimate. Although it is easy to obtain the expected value of the estimate without that interpretation, we shall need it later in the derivation of the variance, and hence we give it here to shorten the total presentation a little.

Suppose that clusters $C_1, \cdots, C_{k_1}$ are such that

$$P_1 + \cdots + P_{k_1-1} < \overline{P}_m \leq P_1 + \cdots + P_{k_1}.$$

Then we define stratum 1 to consist of clusters $C_1, \cdots, C_{k_1}$. It is easy to see that if the above sampling method is used then

$$P\{C_h \text{ is selected from stratum 1}, h < k_1\} = \frac{P_h}{\overline{P}_m},$$

$$P\{C_{k_1} \text{ is selected from stratum 1}\} = \frac{\overline{P}_m - P_1 - \cdots - P_{k_1-1}}{\overline{P}_m}.$$

Furthermore, suppose that clusters $C_{k_1}, \cdots, C_{k_1+k_2}$ are such that

$$P_1 + \cdots + P_{k_1+k_2-1} < 2\,\overline{P}_m \leq P_1 + \cdots + P_{k_1+k_2}.$$

Then we define stratum 2 to consist of clusters $C_{k_1}, \cdots, C_{k_1+k_2}$. It is easy see that if the above sampling method is used, then

$$P\{C_{k_1} \text{ is selected from stratum 2}\} = \frac{P_1 + \cdots + P_{k_1} - \overline{P}_m}{\overline{P}_m},$$

$$P\{C_{k_1+h} \text{ is selected from stratum 2}, 1 \leq h < k_2\} = \frac{P_h}{\overline{P}_m},$$

$$P\{C_{k_1+k_2} \text{ is selected from stratum 2}\} = \frac{2\overline{P}_m - P_1 - \cdots - P_{k_1+k_2-1}}{\overline{P}_m}.$$

Since $P_h \leq \overline{P}_m$ we remark that it is impossible that C_{k_1} be selected from both stratum 1 and stratum 2.

In general, if clusters $C_{k_1+\cdots+k_{i-1}}, \cdots, C_{k_1+\cdots+k_i}$ are such that

$$P_1 + \cdots + P_{k_1+\cdots+k_{i-1}} < i\,\overline{P}_m \leq P_1 + \cdots + P_{k_1+\cdots+k_i} \tag{5.2}$$

then the ith stratum consists of these $k_i + 1$ clusters, and we define the probabilities $P_{i\alpha}$, $\alpha = 0, \cdots, k_i$, by the equations

$$\begin{aligned} P_{i0} &= P\{C_{k_1+\cdots+k_{i-1}} \text{ is selected from stratum } i\} \\ &= \frac{P_1 + \cdots + P_{k_1+\cdots+k_{i-1}} - (i-1)\overline{P}_m}{\overline{P}_m}, \\ P_{i\alpha} &= P\{C_h \text{ is selected from stratum } i, k_1 + \cdots + k_{i-1} < h < k_1 + \cdots + k_i\} \\ &= \frac{P_h}{\overline{P}_m}, \qquad \alpha = h - k, - \cdots - k_{i-1}, \\ P_{ik_i} &= P\{C_{k_1+\cdots+k_i} \text{ is selected from stratum } i\} \\ &= \frac{i\overline{P}_m - P_1 - \cdots - P_{k_1+\cdots+k_{i-1}}}{\overline{P}_m}. \end{aligned} \tag{5.3}$$

We remark that

$$P_{i-1k_{i-1}} + P_{i0} = \frac{P_{k_1+\cdots+k_{i-1}}}{\overline{P}_m}. \tag{5.4}$$

Now, let the elements of the population be x_{hj}, $h = 1, \cdots, M$, $j = 1, \cdots, N_h$, and let the arithmetic mean of the hth cluster be denoted by $\bar{x}_h$. Since the N_h are usually unknown but the measure of size, P_h, is known, we sample, not with probability proportionate to the N_h, but with probability proportionate to the P_h. We shall denote the clusters of the ith stratum by $C_{i0}, \cdots, C_{ik_i}$, making the identification

$$C_{i\alpha} = C_{\alpha+k_1,+\cdots+k_{i-1}}. \tag{5.5}$$

Furthermore, the number of elements of the clusters are denoted by $N_{i0}, \cdots,$

N_{ik_i}, and the means of the clusters by $\bar{x}_{i0}, \cdots, \bar{x}_{ik_i}$, where

$$N_{i\alpha} = N_{\alpha+k_1+\cdots+k_{i-1}}$$

$$\bar{x}_{i\alpha} = \bar{x}_{\alpha+k_1+\cdots+k_{i-1}} \tag{5.6}$$

so that $\bar{x}_{i0} = \bar{x}_{i-1,k_{i-1}}$ and $N_{i0} = N_{i-1,k_{i-1}}$, $i = 1, \cdots, m$.

Furthermore, we define

$$\tilde{x}_{i\alpha} = N_{i\alpha}\,\bar{x}_{i\alpha}/P_{i\alpha} = \tilde{x}_{\alpha+k,+\cdots+k_{i-1}}. \tag{5.7}$$

We define the mean of the ith stratum to be

$$\bar{\bar{x}}_i = \sum_{\alpha=0}^{k_i} P_{i\alpha}\tilde{x}_{i\alpha}/P_m, \tag{5.8}$$

and the variance of the ith stratum to be

$$\sigma_i^2 = \sum_{\alpha=0}^{k_i} \frac{P_{i\alpha}}{P_m}(\tilde{x}_{i\alpha} - \bar{\bar{x}}_i)^2. \tag{5.9}$$

Then, if the mean and variance of the population are defined to be

$$\tilde{x} = \sum_{h=1}^{M} P_h\tilde{x}_h/P \tag{5.10}$$

and

$$\sigma^2 = \sum_{h=1}^{M} \frac{P_h}{P}(\tilde{x}_h - \tilde{x})^2, \tag{5.11}$$

it is easy to verify that

$$\tilde{x} = \frac{1}{m}\sum_{i=1}^{m} \bar{\bar{x}}_i \tag{5.12}$$

and

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m} \sigma_i^2 + \frac{1}{m}\sum_{i=1}^{m} (\bar{\bar{x}}_i - \tilde{x})^2. \tag{5.13}$$

An unbiased estimate of the total of a characteristic. We shall see that we can obtain an estimate of x, where

$$x = \sum_{i=1}^{M}\sum_{j=1}^{N_i} x_{ij}$$

i. e. x is the total of the elements of the population. Since N is unknown, the estimate of $\bar{x}$ that is used is the ratio of unbiased estimates of x and N. It is well known that this ratio is usually biased. Since we are not making any study of ratio estimates here we will not derive the approximation to the variance of this estimate. It may be remarked that it can be obtained by a simple extension of the results here given.

Let us agree that the general form of the estimate will be as follows:

If the jth cluster of the population is selected we shall subsample n_j elements from it. The total of the values of the characteristic for these n_j elements we denote by x_j'. Furthermore, we denote by n_i' the total number of elements subsampled from the ith stratum, or, what is the same, from the cluster selected from the ith stratum; and by x_i'' the total of these elements. Thus, if the jth cluster is the ith selected, then $n_i' = n_j$ and $x_i'' = x_j'$. We define our estimate x'' of x, the total of the population, to be

$$x'' = K(x_1'' + \cdots + x_m''). \tag{5.14}$$

Then, if $K = P/mn$ and $n_h = nN_h/P_h$, it is easy to see that x'' is an unbiased estimate of x.

The variance of the estimate. We may calculate the variance of x'' where

$$x'' = \overline{P}_m(\tilde{x}_1'' + \cdots + \tilde{x}_m'') \quad \text{and} \quad \tilde{x}_i'' = x_i''/n. \tag{5.15}$$

Now, by Theorem 2,

$$\sigma_{x''}^2 = E\sigma_{x''|A}^{2*} + \sigma_{E^*(x''|A)}^2, \tag{5.16}$$

where A^* has been defined above. We shall not evaluate $E\sigma_{x''|A}^{2*}$ since this involves no new problem for subsampling methods using random or systematic methods, or methods using probability proportionate to size.

From (5.15) it follows that

$$E^*(x''|A) = \overline{P}_m(\tilde{x}_1' + \cdots + \tilde{x}_m') \tag{5.17}$$

or, in other words, $E^*(x''|A)$ is the estimate we would have if the clusters in the sample were completely enumerated. We shall denote the second term of (5.16) by σ_B^2. Then,

$$\sigma_B^2 = \overline{P}_m^2\left\{\sum_{i=1}^{m}\sigma_{\tilde{x}_i'}^2 + \sum_{i\neq j}\sigma_{\tilde{x}_i'\tilde{x}_j'}\right\}. \tag{5.18}$$

Now

$$\sigma_{\tilde{x}_i'}^2 = \sum_{\alpha=0}^{k_i}\frac{P_{i\alpha}}{\overline{P}_m}(\tilde{x}_{i\alpha} - \bar{\bar{x}}_i)^2 = \sigma_i^2. \tag{5.19}$$

To calculate $\sigma_{\tilde{x}_i'\tilde{x}_j'}$, $i \neq j$, we shall use Theorem 1.

$$\sigma_{\tilde{x}_i'\tilde{x}_j'} = E(\tilde{x}_i' - \bar{\bar{x}}_i)(\tilde{x}_j' - \bar{\bar{x}}_j) = E\{(\tilde{x}_i' - \bar{\bar{x}}_i)E^*[(\tilde{x}_j' - \bar{\bar{x}}_j)\,|\,\tilde{x}_i']\}. \tag{5.20}$$

To calculate $E^*[(\tilde{x}_j' - \bar{\bar{x}}_j)\,|\,\tilde{x}_i']$ we begin by noting that

$$E^*[(\tilde{x}_j' - \bar{\bar{x}}_j)\,|\,\tilde{x}_i'] \equiv E^*[(\tilde{x}_j' - \bar{\bar{x}}_j)\,|\,C_i] \tag{5.21}$$

where C_1^* is the random event having $k_i + 1$ possible states which are the selections of $C_{i0}, \cdots, C_{ik_i}$ as the sample clusters of the ith stratum. Now if $C_{i\alpha}$ is one of the clusters of the ith stratum let us calculate

$$E[(\tilde{x}_j' - \bar{\bar{x}}_j)\,|\,C_{i\alpha}]. \tag{5.22}$$

We begin by determining which of the clusters of the jth stratum are possible sample clusters, if we know that $C_{i\alpha}$ is selected from the ith stratum. Since the sizes of strata i and j are both $\overline{P}_m$ it follows that there exist integers β_0 and β_1 such that

$$P_{j0} + \cdots + P_{j\beta_0-1} \leq P_{i1} + \cdots + P_{i,\alpha-1} < P_{j1} + \cdots + P_{j\beta_0},$$

and

$$P_{j0} + \cdots + P_{j\beta_1-1} < P_{i1} + \cdots + P_{i\alpha} \leq P_{j1} + \cdots + P_{j\beta_1}.$$

Hence, if we know that $C_{i\alpha}$ has been selected from stratum i, it follows that we must select one of the clusters

$$C_{j\beta_0}, C_{j\beta_0+1}, \cdots, C_{j\beta_1}$$

from stratum j and

$$P\{C_{j\beta} \text{ is selected} \mid C_{i\alpha} \text{ is selected}\} = P'_{j\beta}/P_{i\alpha}, \beta = \beta_0, \beta_0 + 1, \cdots, \beta_1$$
$$= 0, \text{ otherwise,}$$

where

$$P'_{j\beta_0} = P_{j1} + \cdots + P_{j\beta_0} - P_{i1} - \cdots, P_{i,\alpha-1}$$
$$P'_{j\beta} = P_{j\beta}, \beta = \beta_0 + 1, \cdots, B_1 - 1$$
$$P'_{j\beta_1} = P_{i1} + \cdots + P_{i\alpha} - P_{j1} - \cdots - P_{j\beta_1-1},$$

and

$$\sum_{\beta=\beta_0}^{\beta_1} P'_{j\beta} = P_{i\alpha}$$

Then

$$E[(\tilde{x}'_j - \tilde{\bar{x}}_j \mid C_{i\alpha}) = \tilde{x}_{j|\alpha} - \tilde{\bar{x}}_j] \tag{5.23}$$

where

$$\tilde{x}_{j|\alpha} = \sum_{\beta=\beta_0}^{\beta_1} \frac{P'_{j\beta}}{P_{i\alpha}} \tilde{x}_{j\beta}. \tag{5.27}$$

Hence, substituting in (5.20), we see that

$$\sigma_{\tilde{x}'_i\tilde{x}'_j} = E(\tilde{x}'_i - \tilde{\bar{x}}_i)(\tilde{x}'_{j|i} - \tilde{\bar{x}}_j)$$

where $\tilde{x}'_{j|i} = \tilde{x}_{j|\alpha}$ if $C_{i\alpha}$ is selected from stratum i. Then it follows that

$$\sigma_{\tilde{x}'_i\tilde{x}'_j} = \sum_{\alpha=0}^{k_i} \frac{P_{i\alpha}}{\overline{P}_m} (\tilde{x}_{i\alpha} - \tilde{\bar{x}}_i)(\tilde{x}_{j|\alpha} - \tilde{\bar{x}}_j). \tag{5.25}$$

Obviously, the conditional expectation can be eliminated from (5.25) by using (5.23) but no gain in simplicity or generality thus occurs.

It would be possible to obtain the variances and covariances of the x_i' by listing all possible samples in any special case. To make this general would only require writing the necessary notation.

Substituting in (5.18) we see that

$$\sigma_B^2 = \overline{P}_m^2 \left\{ \sum_{i=1}^{m} \sigma_i^2 + \sum_{i \neq j} \sigma_{z_i' z_j'} \right\}$$

where $\sigma_{z_i' z_j}$ is given by (5.25).

It follows that if we use the fact that $\sum_{i=1}^{m} (\bar{\bar{x}}_i - \bar{x}) = 0$, then we have

$$\sigma_B^2 = \overline{P}_m^2 \sum_{i=1}^{m} \sum_{\alpha=0}^{k_i} \frac{P_{i\alpha}}{\overline{P}_m} (\bar{x}_{i\alpha} - \bar{x})^2 + \overline{P}_m^2 \sum_{i \neq j} \sum_{\alpha=0}^{k_i} \frac{P_{i\alpha}}{\overline{P}_m} (\bar{x}_{i\alpha} - \bar{x})(\bar{x}_{j|\alpha} - \bar{x}),$$

or, returning in part to the "unstratified" notation

$$(5.26) \qquad \sigma_B^2 = \frac{P^2}{m} \sum_{h=1}^{M} \frac{P_h}{P} (\bar{x}_h - \bar{x})^2 + \frac{P^2}{m} \sum_{i \neq j} \sum_{\alpha=0}^{k_i} \frac{P_{i\alpha}}{P} (\bar{x}_{i\alpha} - \bar{x})(\bar{x}_{j|\alpha} - \bar{x}).$$

By combining terms of the second part of (4.26) generalizations of the formulae obtained in [1] are easily obtained.

Still another means of writing σ_B^2 is

$$(5.27) \qquad \sigma_B^2 = \frac{P^2}{m} \left\{ \sigma^2 - \sigma_{b \cdot s \cdot}^2 + \sum_{i \neq j} \sum_{\alpha=0}^{k_i} \frac{P_{i\alpha}}{P} (\bar{x}_{i\alpha} - \bar{x})(\bar{x}_{j|\alpha} - \bar{x}) \right\}$$

where

$$\sigma_{b \cdot s \cdot}^2 = \frac{1}{m} \sum_{i=1}^{m} (\bar{\bar{x}}_i - \bar{x})^2,$$

which shows both sources of changes in efficiency as compared with sampling with probability proportionate to size, and replacing the clusters obtained. (It is, of course, obvious that $P^2\sigma^2/m$ is the variance of $E^*(x' \mid A)$, if we assume the m clusters to have been selected with probability proportionate to size, each selected cluster being replaced before the next is selected.)

By considering (5.26) and (5.27) it is clear that systematic sampling with probability proportionate to size will be more efficient than sampling p.p.s. with replacement under much the same conditions as when we sample single elements. The details are omitted. They depend on applying the Lemma and Theorem 4. The summary of the conditions is: If we sample systematically with p.p.s., and if the two sets $x_{i1}, \cdots, x_{ik_i}$ and $y_{ji}, \cdots y_{jk_j}$ are monotone, one being monotone non-icreasing and the other monotone non-decreasing, then the covariance between the ith and jth strata will be negative, and thus gains made as compared with independent sampling from the strata.

If we define

$$\sigma_{\alpha\beta}^0 = \mathfrak{E}(\tilde{x}_{i\alpha} - \mathfrak{E}\tilde{x}_{i\alpha})(\tilde{x}_{j\beta} - \mathfrak{E}\tilde{x}_{j\beta})$$

then the concavity condition for systematic sampling p.p.s. to yield a smaller variance than independent sampling p.p.s. from each stratum is, if $\alpha < \beta$,

$$\sigma^0_{\alpha 1} - \sigma^0_{\gamma 1} \leq \sigma^0_{\alpha 2} - \sigma^0_{\gamma 2} \leq \cdots \leq \sigma^0_{\alpha k_j} - \sigma^0_{\gamma k_j} \leq 0.$$

6. The systematic sampling of clusters of equal sizes. Let us now suppose that our population consists of clusters of elements, the clusters being of equal size, i.e. containing the same number of elements. To be specific, let the population consist of M clusters, where $M = cm$ and each cluster contains N elements, where $N = kn$. Then, the value of the characteristic being measured for the αth element of the ith cluster may be denoted by $x_{i\alpha}$, and the total of all the elements of the ith cluster may be denoted by x_i. The arithmetic mean of the population is $\bar{x}$, and thus

$$M\bar{x} = \sum_{i=1}^{M} \bar{x}_i.$$

where

$$N_i\bar{x}_{i\cdot} = x_i .$$

a. *Complete enumeration of clusters in sample.* First, suppose that we wish to estimate $\bar{x}$ by $\bar{x}'$, where $\bar{x}'$ is the arithmetic mean of the sample obtained by selecting a systematic sample of m of the M clusters, and enumerating all elements within each cluster in the sample. Then, we may write

$$m\bar{x}' = \sum_{i=1}^{m} \bar{x}'_i, \tag{6.1}$$

where $\bar{x}'_i$ is the mean of the ith cluster selected for the sample. From [1], it follows then that

$$\sigma^2_{\bar{x}'} = \frac{\sigma^2_b}{m}\{1 + (m-1)\bar{\rho}_c\}$$

where $M\sigma^2_b = \sum_{i=1}^{M} (\bar{x}_{i\cdot} - \bar{x})^2$, and $\bar{\rho}_c$ is defined as $\bar{\rho}_k$ in [1, p. 6], but with $\bar{x}_i$ in place of x_i. Now from the theory of the random sampling of clusters it follows that

$$\sigma^2_b = \frac{\sigma^2}{N}\{1 + (N-1)\rho\}$$

where σ^2 is the variance of the population, i. e.

$$MN\sigma^2 = \sum_{i=1}^{M}\sum_{j=1}^{N} (x_{ij} - \bar{x})^2$$

and ρ is the intraclass correlation coefficient of elements within clusters, i. e.

$$\sigma^2\rho = \sigma^2_b - \sigma^2_w/N - 1,$$

where

$$MN\sigma_w^2 = \sum_{i=1}^{M}\sum_{j=1}^{N}(x_{ij} - \bar{x}_i)^2.$$

Thus

$$\sigma_{\bar{x}'}^2 = \frac{\sigma^2}{mN}\{1 + (N-1)\rho\}\{1 + (m-1)\bar{\rho}_c\}. \tag{6.2}$$

Of the three factors in (6.2), σ^2/mN is the variance of a random sample of size mN selected with replacement; $1 + (N-1)\rho$ is the factor arising from the use of clusters; and $1 + (m-1)\bar{\rho}_c$ is the factor arising from the fact that the clusters are sampled systematically.

b. *Stratification and subsampling.* When we consider the possibilities of stratification and subsampling, the number of possible designs increases tremendously. For example, it would be simple to calculate the variances of arithmetic means obtained by stratifying the population, selecting sampling units with probability proportionate to size, subsampling systematically, again subsampling systematically and finally subsampling at random. However, such studies may be left to be made in connection with the practical problems in which they are to be used. Rather than attempt to consider many of the possibilities that might arise in practice, we shall here give only the results of the systematic subsampling of a systematic sample. The variances of many other designs may easily be obtained by means of Theorems 1 and 2.

Suppose now that from each of a systematically selected sample of m clusters we subsample, systematically, n elements. Then, let our estimate of $\bar{x}$ be $\bar{x}'$ where, if $x'_{i\alpha}$ is the αth selected element from the ith sample cluster, then

$$\bar{x}' = \left(\frac{1}{mn}\right)\sum_{i=1}^{m}\sum_{\alpha=1}^{n}x'_{i\alpha} = \frac{1}{m}\sum_{i=1}^{m}\bar{x}''_i$$

and

$$\bar{x}''_i = \left(\frac{1}{n}\right)\sum_{\alpha=1}^{m}x'_{i\alpha}.$$

From Theorem 2, it follows at once that

$$\sigma_{\bar{x}'}^2 = \frac{\sigma^2}{mN}\{1 + (N-1)\rho\}\{1 + (m-1)\bar{\rho}_c\} + \frac{1}{M}\sum_{i=1}^{M}\frac{\sigma_i^2}{mn}\{1 + (m-1)\bar{\rho}_i\}, \tag{6.3}$$

where σ_i^2 is the variance within the ith cluster and ρ_i is the average serial correlation within the ith cluster as defined in [1, p. 6]. It is simple to calculate the variance of $\bar{x}'$ also when the sub-sampling is done by considering the m clusters in the sample as one population from which a systematic sample is selected. This is the case that occurs when a sample of blocks is selected and all the households on the sample blocks are listed serially, a systematic sample then being selected from the lists. However, for our present purposes it is the analysis of (6.2) that is important and we now turn to a brief discussion of (6.2).

The most important conclusion to be drawn from (6.2) is that the systematic

selection of clusters even when systematic selection is desirable, may not compensate for the increase in variance caused by the use of clusters. Systematic selection will provide the same relative gains but these gains may not be large enough to produce the inequality

$$\{1 + (N - 1)\rho\}\{1 + (m - 1)\bar{\rho}_c\} < \frac{MN - mN}{MN - 1}.$$

A problem that we have not worked through is the following: By regarding the elements of the population as random variables, we obtain conditions on the average correlations among elements of a single cluster as well as on the average correlations among elements of different clusters that enable us to state where the systematic sampling of clusters of equal sizes may be expected to yield a smaller variance than the random or stratified random sampling of clusters or of individual elements. This solution should be straight forward.

c. *Systematic sampling in two dimensions.* Systematic sampling in two dimensions occurs in such practical problems as the selection of a sample of blocks from a city or the selection of a sample of plots from a field.

In selecting blocks from a city, the procedure most often followed effectively reduces the problem to one dimensional form by first numbering the blocks of the city or a part of it, in serpentine fashion beginning, say, in the upper right corner of a map of the city and numbering the blocks in the top row from right to left continuing the numbering of the second row from left to right and so on. Then a systematic sample of these block numbers, and hence, of the blocks themselves is selected. Clearly, this procedure should not be the most efficient if neighboring blocks are highly correlated, since, to cite an unrealistic possibility, the possible samples might turn out to be columns of blocks of the city.

A second two dimensional systematic sampling procedure might be that of selecting a systematic sample of the rows and a systematic sample of the columns, thus obtaining a grid sample. This design too is inefficient when there is a "fertility gradient" along rows or along columns.

The reason for the inefficiency of both of these procedures can be found by examining the formulae for the variances of systematic samples. If the numbering is serpentine, then it becomes illogical to expect that the correlogram is concave upwards and sharp deviations from that pattern may occur. In the grid design, which is a special case of the systematic sampling of clusters with systematic subsampling, we may examine (6.3) and note that the intra-class correlation coefficient ρ may be large enough for $\sigma^2_{\bar{x}'}$ to be large even when $\bar{\rho}_c$ is negative.

Clearly, (6.3) suggests that the possible samples be so defined that ρ is as small as possible. In square fields this might be attained by defining the possible samples to be plots of a Knut Vik square having the same treatment, and similar definitions of possible samples could easily be given for irregular fields. This subject is, however, left for further study.[3]

---

[3] One of the referees of this paper has drawn the author's attention to an article [6], the data of which, especially Table 3, are in accordance with the opinions expressed above.

REFERENCES

[1] W. G. Madow and L. H. Madow, "On the theory of systematic sampling, I," *Annals of Math. Stat.*, Vol. 15 (1944), pp. 1–24.

[2] L. H. Madow, "Systematic sampling and its relation to other sampling designs," *Am. Stat. Asso. Jour.*, Vol. 41 (1946), pp. 204–217.

[3] W. G. Cochran, "Relative accuracy of systematic and stratified random samples for a certain class of populations," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 164–177.

[4] G. H. Hardy, J. E. Littlewood and G. Polya, *Inequalities*, Cambridge University Press, London and New York, 1934, p. 43.

[5] M. H. Hansen and W. N. Hurwitz, "On the theory of sampling from finite populations," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 333–362.

[6] P. G. Homeyer and C. A. Black, "Sampling replicated field experiments on oats for yield determinations," *Socl. Sci. Soc. Proc.*, Vol. 11 (1946), pp. 341–344.

PROBLEMS IN PLANE SAMPLING

By M. H. Quenouille

Rothamsted Experimental Station, Harpenden, England

1. Summary. After consideration of the relative accuracies of systematic and stratified random sampling in one dimension the problem of estimation of linear sampling error is discussed.

Methods of sampling an area are proposed, and expressions for the accuracies of these methods are derived. These expressions are compared for large samples, with special reference to correlation functions which appear to be theoretically and practically justified, and systematic sampling is found to be more accurate than stratified random sampling in many cases. Methods of estimating sampling errors are again considered, and examples given. The paper concludes with some remarks on the problem of trend in the population sampled.

2. Accuracy of systematic and stratified random samples in one dimension. W. G. Cochran [1] has given expressions to the variances of the means of samples of size n drawn from a population $x_1x_2 \cdots x_{nk}$ when the method of sampling is random (r), stratified random (st) and systematic (sy). He assumes the elements $x_1x_2 \cdots x_{nk}$ to be drawn from a population in which

$$E(x_i) = \mu, \qquad E(x_i - \mu)^2 = \sigma^2, \qquad E(x_i - \mu)\,(x_{i+u} - \mu) = \rho_u\sigma^2$$

where $\rho_u \geq \rho_v \geq 0$ whenever $u < v$, and derives the expressions

$$(1) \qquad \sigma_r^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\left[1 - \frac{2}{kn(kn-1)}\sum_{u=1}^{kn-1}(kn-u)\rho_u\right]$$

$$(2) \qquad \sigma_{st}^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right)\left[1 - \frac{2}{k(k-1)}\sum_{u=1}^{k-1}(k-u)\rho_u\right]$$

$$(3) \qquad \sigma_{sy}^2 = \frac{\sigma^2}{n}\left(1 - \frac{1}{k}\right) \cdot \left[1 - \frac{2}{kn(k-1)}\sum_{u=1}^{kn-1}(kn-u)\rho_u + \frac{2k}{n(k-1)}\sum_{u=1}^{n-1}(n-u)\rho_{ku}\right].$$

Using these expressions which are linear functions of the ρ_u Cochran compares the relative efficiencies of the methods of sampling for several types of correlogram. It is worth noting that (1), (2) and (3) can be derived under more general conditions than Cochran considered. If we assume that (a) each x_i is a sample from a population with mean μ_i and variance σ_i^2, (b) that μ_i is distributed about mean μ with variance σ^2, (c) that $E(\mu_i - \mu)\,(\mu_j - \mu) = \rho_{ij}\sigma^2$, and (d) that $\rho_u = \dfrac{1}{kn-u}\displaystyle\sum_{i=1}^{kn-u}\rho_{i,i+u}$, then it is not difficult to show that (1), (2) and (3)

require the addition of a superposed variation $\frac{1}{n}\left(1-\frac{1}{k}\right)\cdot\frac{1}{kn}\sum_{i=1}^{kn}\sigma_i^2$ to the right-hand side of the equations. Thus it should be remembered that Cochran's results give theoretical maxima to the relative efficiencies of the various methods of sampling, while ρ_u is the mean correlation between samples u apart. This result is perhaps interesting in connection with sampling for say, insect infestation, when at each point there will be a mean level of infestation and the sample will be distributed in a Poisson distribution about this mean. Then the superposed variation is

$$\frac{1}{n}\left(1-\frac{1}{k}\right)\cdot\frac{1}{kn}\sum_{i=1}^{kn}\mu_i\sim\frac{1}{n}\left(1-\frac{1}{k}\right)\mu.$$

If we are sampling a continuous process[1], for n large we can write down the integral equivalents of (1), (2) and (3)

$$\sigma_r^2\sim\frac{\sigma^2}{n}$$

$$\sigma_{st}^2\sim\frac{\sigma^2}{n}\left[1-\frac{2}{d^2}\int_0^d(d-u)\rho_u\delta u\right] \tag{4}$$

$$\sigma_{sy}^2\sim\frac{\sigma^2}{n}\left[1-\frac{2}{d}\int_0^\infty\rho_u\delta u+2\sum_{u=1}^{\infty}\rho_{du}\right] \tag{5}$$

where ρ_u is the mean correlation between successive elements of the sample, u apart and d is the mean distance between samples. We have thus

$$\frac{\sigma_{st}^2-\sigma_{sy}^2}{\sigma_r^2}\sim\frac{2}{d}\left[\int_0^d\frac{u}{d}\rho_u\delta u+\int_d^\infty\rho_u\delta u-d\sum_{u=1}^{\infty}\rho_{du}\right],$$

which can often be used to investigate, quickly and roughly, with the aid of a graph the difference between the efficiencies of stratified random and systematic sampling. Figure 1 shows how this is done for four types of correlogram.

For a continuous Markoff scheme, we have $\rho_u=\rho^u$ and

$$\sigma_{st}^2\sim\frac{\sigma^2}{n}\left[1+\frac{2}{\log\rho^d}+\frac{2}{(\log\rho^d)^2}-\frac{2\rho^d}{(\log\rho^d)^2}\right],$$

$$\sigma_{sy}^2\sim\frac{\sigma^2}{n}\left[1+\frac{2}{\log\rho^d}+\frac{2\rho^d}{1-\rho^d}\right],$$

which agree with Cochran's results.

3. Replication and the estimation of error. Yates [2] has pointed out the difficulties attached to the estimation of error for a systematic sample. It will, however, be worthwhile to investigate this point using the above formulae.

[1] In practice we can sample a continuous process only as if it were a discontinuous process with k large.

For random, stratified random and systematic sampling, if n is large and k is regarded as constant, then the variance of the estimate of the mean will be of the form $\sigma^2 F(k)/n$, where $F(k)$ is virtually independent of n. Thus, if we have any method which provides an estimate of error for the samples it will be possible to split the series to be sampled into several equal parts (or blocks) to obtain an estimate of error of the mean of each part and to combine these to obtain a more accurate estimate of the error of the overall mean. In fact, if n is very large, we may wish to reduce our number of observations by obtaining estimates of error from a random selection of these parts. For stratified random sampling, $F(k)$ is completely independent of n, so that we may combine our estimates of error from each strata. This leads us to the commonly used method of taking q randomly chosen elements per strata, and combining the sets of variances of $q - 1$ degrees of freedom to form an estimate of error. If we make our samples exclusive, i.e. no two elements can coincide, then this variance has to be multiplied by $1 - q/k$ to give the estimated variance of the sample mean.

We can in the same way estimate the variance of the mean of a systematic sample by using sets of q systematic samples of sufficient length with randomly-chosen starting points. This sampling will, however, be more difficult to carry out in practice, and we might consider other methods. Our systematic samples may be chosen to be invariable in each part or block into which the series is split so that our sampling procedure involves, in all, only q systematic samples, or we might follow the method advocated by Yates of choosing our q samples to be evenly spaced, so that they are subsamples of a larger systematic sample. Whereas this latter method has simplicity and its possible incorporation into a more extensive scheme to recommend it, its use has to be very carefully considered. If we consider the discrete case, we wish to estimate

$$\sigma^2 \left(1 - \frac{2}{k-1} \sum_{u=1}^{\infty} \rho_u + \frac{2k}{k-1} \sum_{u=1}^{\infty} \rho_{ku}\right), \tag{6}$$

but any estimate of variance based on q evenly-spaced systematic samples can contain only terms of the form $\rho_{ku/q}$, and while an estimate of variance based on q randomly-chosen systematic samples will obviously be limited, it will, in most cases, be more representative. As an example, suppose we take $k = 16$ and $q = 4$ then we can compare the relative occurrences of observing the correlations $\rho_1 \cdots \rho_{15}$ in the estimate of variance. Six examples of this are given in table 1, the random numbers having been drawn from Fisher and Yates tables; ρ_u and ρ_{16-u} being shown together, since they occur equally frequently. The table demonstrates how randomly-chosen samples, even as nearly systematic as the first two randomly-chosen samples will avoid systematically sampling the correlogram. It is obvious that in most cases either method will be fairly good but the use of this latter will usually be the more accurate. Comparisons are made in table 2 for various types of correlogram using the samples indicated in table 1. It is, of course, possible to postulate theoretically many kinds of

[2] Throughout this paper δ is used for the differential sign to prevent confusion with d.

correlogram for which the equal-spaced sets of systematic samples will break down, but ultimately we must decide with reference to the types of correlogram

TABLE 1

Frequency of occurrence of the serial correlations ρ_1 , $\rho_2 \ldots \rho_{15}$ *in the estimate of variance when 4 systematic samples each with spacing 16 units are taken*

ρ	4 evenly-spaced systematic samples	4 systematic samples with random starting points at						Total frequencies
		4, 7, 8, 12,	3, 7, 8, 12,	3, 6, 10, 13,	4, 6, 7, 14,	2, 8, 11, 15,	2, 6, 11, 16	
1, 15		1	1		1			3
2, 14					1		1	2
3, 13		1		2	1	2		6
4, 12	4	2	2	1		1	1	7
5, 11		1	2				2	5
6, 10				1	1	1	1	4
7, 9			1	2	1	2	1	7
8	4	2			2			4

TABLE 2

Values of $\frac{2}{15}\sum_{u=1}^{15}\rho_u$ *as estimated by systematic samples*

ρ_u	Evenly-spaced systematic samples	Systematic samples with random starting points						Mean	Expected
		1	2	3	4	5	6		
$1-0.2\,u$, $(u = 1, .5)$	0.17	0.27	0.20	0.17	0.30	0.17	0.13	0.21	0.27
$1-0.1\,u$, $(u = 1, .10)$	0.53	0.62	0.58	0.53	0.60	0.53	0.53	0.57	0.60
2^{-u}	0.04	0.13	0.12	0.06	0.15	0.06	0.07	0.10	0.13
$2^{-u/4}$	0.58	0.66	0.64	0.60	0.66	0.60	0.60	0.63	0.65
Kendall's Series 1	−0.14	0.03	0.00	−0.05	0.16	−0.05	−0.05	0.01	0.07

* Naturally the use of this method of estimating the sampling error assumes that the correlation between the corresponding elements in each part or block into which the series is split may be neglected, i.e. in this case that the terms ρ_{16} and above are negligible. In this case $\rho_{16} = 1/16$ and consequently the term $2(\frac{1}{15}\sum_{u=1}^{\infty}\rho_u - \frac{16}{15}\sum_{u=1}^{\infty}\rho_{16u}) = 0.56$, required in (6) differs slightly from the term $\frac{2}{15}\sum_{u=1}^{15}\rho_u = 0.65$ which we are attempting to estimate.

experienced. We shall consider this point further, after we have dealt with two-dimensional sampling.

4. Methods of sampling in 2 dimensions. The number of ways in which we can sample a two-dimensional space[3] is large, since we can employ random,

[3] We shall, in general, consider our two-dimensional space to be rectangular, but it is not difficult to draw similar conclusions for an area of any shape.

stratified random or systematic sampling in either direction. Thus we will be able to consider every possible combination of these methods, e.g. random in

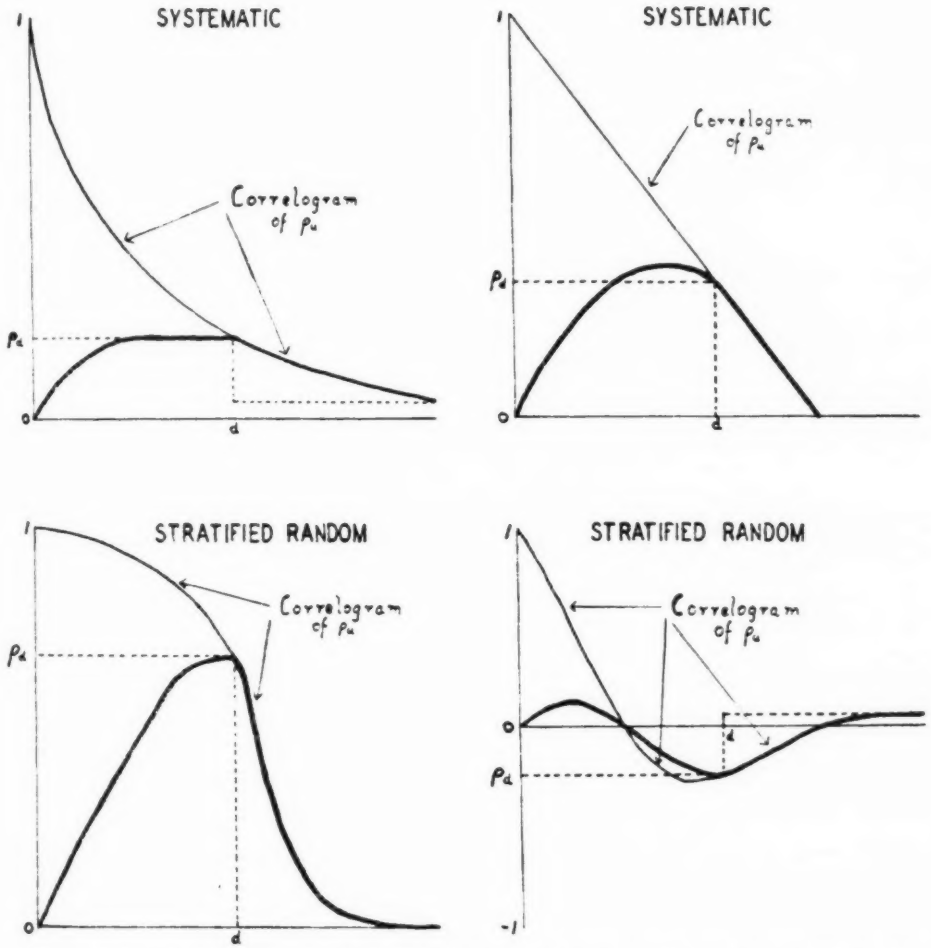

FIG. 1. Graphical comparison of the efficiencies of systematic and stratified random sampling for various correlation functions. The thick line gives the function

$$f_1(u) = u\rho_u/d, \qquad 0 \leq u \leq d$$
$$= \rho_u, \qquad d \leq u,$$

and the dotted line the function

$$f_2(u) = \rho_{id}, \qquad (i-1)d < u \leq id.$$

Thus systematic sampling is more or less efficient than stratified random sampling according to whether the area under the thick line is greater or less than the area under the dotted line. The most efficient method is indicated on each graph.

one direction and systematic in another will be denoted by $r \cdot sy$. Furthermore we can consider the sets of samples in one direction to be aligned with one another, or to be independently determined. The suffix 1 will be used to denote

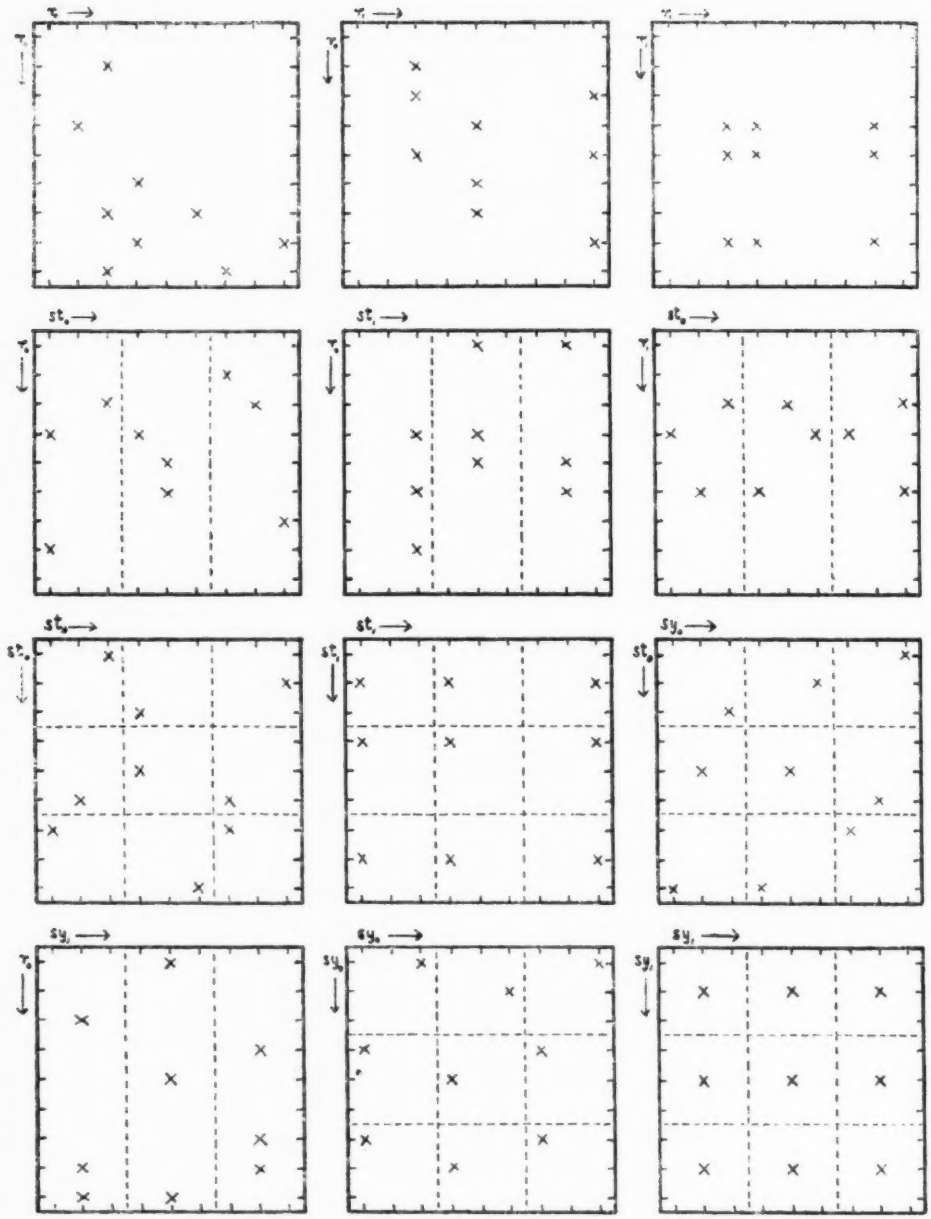

FIG. 2. Methods of sampling a field. In this case, $n_1 = n_2 = k_1 = k_2 = 3$.

aligned samples while suffix 0 will denote independent samples, e.g. we might sample according to the system $r_1 s y_0$. Examples of several methods of sampling are given in Figure 2.

5. Accuracy of sampling in two dimensions. Suppose we consider a sample of n_1n_2 elements drawn from the elements $x_{ij}(i = 1, 2, \cdots n_1k_1, j = 1, 2, \cdots n_2k_2)$, (which form a single finite population drawn from an infinite hypothetical population), such that the mean spacing in the two directions is k_1 and k_2. These parameters will, if necessary, be indicated in brackets after the method of sampling, e.g. $r_1sy_0(n_1k_1\,;\, n_2k_2)$.

Let X denote the mean of a sample formed by the method considered, and x' a member of this sample. Suppose, also, that the x_{ij} are drawn from a population in which

$$E(x_{ij}) = \mu, \qquad E(x_{ij} - \mu)^2 = \sigma^2,$$
$$E(x_{ij} - \mu)\,(x_{i+u,j+v} - \mu) = \rho_{ijuv}\sigma^2,$$

Further we may average ρ_{ijuv} over all possible values of i and j to define $\rho_{uv} = \rho_{-u,-v}$ by the relation

$$\sum_i \sum_j \rho_{ijuv} = (k_1n_1 - |\,u\,|)(k_2n_2 - |\,v\,|)\rho_{uv}.$$

The purpose of these definitions is to allow to eliminate the difficulties associated with the parameters of finite populations by considering this population as being itself a sample from an infinite population. Cochran employs a similar device.

5a. *Random sampling.* It is not difficult to see that

$$\sigma^2(X) = \tfrac{1}{2}\,E(X_1 - X_2)^2 = E(X_1 - \mu)^2 - E(X_1 - \mu)\,(X_2 - \mu),$$

where X_1 and X_2 are independent samples.
Also

$$E(X_1 - \mu)(X_2 - \mu) = E(x_1' - \mu)(x_2' - \mu)$$
$$= \frac{\sigma^2}{k_1k_2n_1n_2}\left[1 + \frac{1}{k_1k_2n_1n_2}\sum\sum (k_1n_1 - |\,u\,|)(k_2n_2 - |\,v\,|)\rho_{uv}\right]$$

where the double summation[4] exists over the region S given by $|\,u\,| \leq k_1n_1$, $|\,v\,| \leq k_2n_2$ and excludes $u = v = 0$. We thus have to evaluate $E(X_1 - \mu)^2$ for the different types of random sampling.

It is easily shown that

$$E(X_1 - \mu)^2 = \frac{\sigma^2}{n_1n_2}$$
$$\cdot\left[1 + \frac{n_1n_2 - 1}{k_1k_2n_1n_2(k_1k_2n_1n_2 - 1)}\sum\sum (k_1n_1 - |\,u\,|)(k_2n_2 - |\,v\,|)\rho_{uv}\right]$$

for r_0r_v,

$$= \frac{\sigma^2}{n_1n_2}\left[1 + \frac{n_1 - 1}{k_1k_2n_1(k_1k_2n_1n_2 - 1)}\sum\sum (k_1n_1 - |\,u\,|)(k_2n_2 - |\,v\,|)\rho_{uv}\right.$$
$$\left. + \frac{2(n_2 - 1)}{k_2n_2(k_2n_2 - 1)}\sum_{v=1}^{k_2n_2}(k_2n_2 - v)\rho_{0v}\right]$$

[4] In general, unless otherwise stated, double summations will exist over the region for which the coefficients are positive, excluding $u = v = 0$.

for $r_1 r_0$,

$$= \frac{\sigma^2}{n_1 n_2}\left[1 + \frac{(n_1-1)(n_2-1)}{k_1 k_2 n_1 n_2(k_1 k_2 n_1 n_2 - 1)} \sum\sum (k_1 n_1 - |u|)(k_2 n_2 - |v|)\rho_{uv}\right.$$

$$\left. + \frac{2(n_2-1)}{k_2 n_2(k_2 n_2 - 1)} \sum_{v=1}^{k_2 n_2} (k_2 n_2 - v)\rho_{0v} + \frac{2(n_1-1)}{k_1 n_1(k_1 n_1 - 1)} \sum_{u=1}^{k_1 n_1} (k_1 n_1 - u)\rho_{u0}\right]$$

for $r_1 r_1$,

whence

$$\sigma^2(r_0 r_0) = \frac{1}{n_1 n_2}\left(1 - \frac{1}{k_1 k_2}\right)\sigma^2$$

$$(7) \qquad \cdot\left[1 - \frac{1}{k_1 k_2 n_1 n_2(k_1 k_2 n_1 n_2 - 1)} \sum\sum (k_1 n_1 - |u|)(k_2 n_2 - |v|)\rho_{uv}\right]$$

$$\sigma^2(r_1 r_0) = \frac{1}{n_1 n_2}\left(1 - \frac{1}{k_1 k_2}\right)\sigma^2\left[1 - \frac{k_1 k_2 n_2 - 1}{(k_1 k_2 - 1)k_1 k_2 n_1 n_2(k_1 k_2 n_1 n_2 - 1)}\right.$$

$$(8) \qquad \cdot\sum\sum (k_1 n_1 - |u|)(k_2 n_2 - |v|)\rho_{uv}$$

$$\left. + \frac{2(n_2-1)}{k_2 n_2(k_2 n_2 - 1)} \sum_{1}^{k_2 n_2} (k_2 n_2 - v)\rho_{0v}\right]$$

$$\sigma^2(r_1 r_1) = \frac{1}{n_1 n_2}\left(1 - \frac{1}{k_1 k_2}\right)\sigma^2\left[1 - \frac{k_1 k_2(n_1 + n_2 - 1) - 1}{(k_1 k_2 - 1)k_1 k_2 n_1 n_2(k_1 k_2 n_1 n_2 - 1)}\right.$$

$$(9) \qquad \cdot\sum\sum (k_1 n_1 - |u|)(k_2 n_2 - |v|)\rho_{uv} + \frac{2k_1(n_2-1)}{(k_1 k_2 - 1)n_2(k_2 n_2 - 1)}$$

$$\left. \cdot \sum_{v=1}^{k_2 n_2} (k_2 n_2 - v)\rho_{ov} + \frac{2k_2(n_1-1)}{(k_1 k_2 - 1)n_1(k_1 n_1 - 1)} \sum_{u=1}^{k_1 n_1} (k_1 n_1 - u)\rho_{u0}\right].$$

5b. *Stratified random sampling.* We can deduce the variances for some methods of taking stratified samples if $\bar{x}_i'$, the mean of the elements sampled in the ith stratum, is independent of $\bar{x}_j'$, since we will then have

$$E(X \dot{-} \bar{x})^2 = E(\bar{x}_i' - \bar{x}_i)^2/n,$$

where $\bar{x}$ is the mean of the finite population which is sampled. Hence

$$\sigma^2(st_0 r_0) = \frac{1}{n_1}\sigma^2\{r_0 r_0(1, k_1 ; n_2 k_2)\}$$

$$(10) \qquad = \frac{1}{n_1 n_2}\left(1 - \frac{1}{k_1 k_2}\right)\sigma^2\left[1 - \frac{1}{k_1 k_2 n_2(k_1 k_2 n_2 - 1)}\right.$$

$$\left. \cdot\sum\sum (k_1 - |u|)(k_2 n_2 - |v|)\rho_{uv}\right].$$

$$\sigma^2(st_1\, r_0) = \frac{1}{n_1 n_2}\left(1 - \frac{1}{k_1 k_2}\right)\sigma^2\left[1 - \frac{1}{k_1 k_2 n_2(k_1 k_2 - 1)}\right.$$

$$\cdot \sum\sum (k_1 - |\, u\,|)(k_2 n_2 - |\, v\,|)\rho_{uv}$$

(11)

$$\left. + \frac{2k_1(n_2 - 1)}{(k_1 k_2 - 1)n_2(k_2 n_2 - 1)} \sum_{v=1}^{k_2 n_2} (k_2 n_2 - v)\rho_{uv}\right],$$

$$\sigma^2(st_0\, st_0) = \frac{1}{n_1 n_2}\left(1 - \frac{1}{k_1 k_2}\right)\sigma^2\left[1 - \frac{1}{k_1 k_2(k_1 k_2 - 1)}\right.$$

(12)

$$\left. \cdot \sum\sum (k_1 - |\, u\,|)(k_2 - |\, v\,|)\rho_{uv}\right].$$

To estimate the variance of other methods of sampling, we will make use of a general formula which we might have used to derive the expressions (8)–(12).

If x_i' is any element of the sample X, then

$$(X - \bar{x})^2 = \frac{1}{n_1 n_2}\left[\sum (x_i' - \bar{x})^2 - \sum (x_i' - X)^2\right]$$

$$= \frac{1}{n_1 n_2}\left[\sum (x_i' - \bar{x})^2 - \frac{n_1 n_2 - 1}{n_1 n_2}\sum (x_i' - \mu)^2\right.$$

$$\left. + \frac{2}{n_1 n_2}\sum\sum (x_i' - \mu)(x_j' - \mu)^{\delta}\right],$$

whence

$$\sigma^2(X) = E(X - \bar{x})^2$$

$$= \frac{k_1 k_2 n_1 n_2 - 1}{k_1 k_2 n_1 n_2}\sigma^2\left[1 - \frac{1}{k_1 k_2 n_1 n_2(k_1 k_2 n_1 n_2 - 1)}\sum\sum (k_1 n_1 - |\, u\,|)\right.$$

(13)

$$\left. \cdot (k_2 n_2 - |\, v\,|)\rho_{uv}\right] - \frac{n_1 n_2 - 1}{n_1 n_2}\sigma^2 + \frac{n_1 n_2 - 1}{n_1 n_2} E(x_i' - \mu)(x_j' - \mu)$$

$$= \frac{1}{n_1 n_2}\left(1 - \frac{1}{k_1 k_2}\right)\sigma^2\left[1 - \frac{1}{k_1 k_2 n_1 n_2(k_1 k_2 - 1)}\sum\sum (k_1 n_1 - |\, u\,|)\right.$$

$$\left. \cdot (k_2 n_2 - |\, v\,|)\rho_{uv} + \frac{k_1 k_2(n_1 n_2 - 1)}{k_1 k_2 - 1}\frac{E(x_i' - \mu)(x_j' - \mu)}{\sigma^2}\right].$$

Thus, provided that we can estimate $E(x_i' - \mu)\,(x_j' - \mu)/\sigma^2$ the expression (13) gives the error for all methods of sampling.

As an example, we might deduce the expression (12). If we choose any member x_i', then a second member x_j' will be located at random with respect to x_i' except that there will be $k_1 k_2 - 1$ positions in the same stratum as x_i' that x_j' will not be able to occupy. Thus the expected correlation $E(x_i' - \mu)\,(x_j' - \mu)/\sigma^2$ will be given by

$$\frac{1}{k_1^2 k_2^2 n_1 n_2(n_1 n_2 - 1)}\sum\sum (k_1 n_1 - |\, u\,|)(k_2 n_2 - |\, v\,|)\rho_{uv}$$

(14)

$$- \frac{1}{k_1^2 k_2^2(n_1 n_2 - 1)}\sum\sum (k_1 - |\, u\,|)(k_2 - |\, v\,|)\rho_{uv}.$$

If we substitute (14) into (13), we will obtain expression (12) for the variance of st_0st_0. In the same manner, we can derive for st_1st_1 the expression

$$
\begin{aligned}
&E(x_i' - \mu)(x_j' - \mu) \\
&\quad = \frac{1}{k_1 k_2(n_1 n_2 - 1)}\left[\frac{1'}{k_1 k_2 n_1 n_2}\sum\sum (k_1 n_1 - |\, u\,|)(k_2 n_2 - |\, v\,|)\rho_{uv}\right. \\
&\qquad - \frac{1}{k_1 k_2 n_1}\sum\sum (k_1 n_1 - |\, u\,|)(k_2 - |\, v\,|)\rho_{uv} \\
&\qquad - \frac{1}{k_1 k_2 n_2}\sum\sum (k_1 - |\, u\,|)(k_2 n_2 - |\, v\,|)\rho_{uv} + \frac{1}{k_1 k_2}\sum\sum (k_1 - |\, u\,|) \\
&\qquad \cdot (k_2 - |\, v\,|)\rho_{uv} + \frac{2(k_1 k_2 n_1 - 1)}{k_1 n_1(k_1 n_1 - 1)}\sum_{u=1}^{k_1 n_1}(k_1 n_1 - u)\rho_{0v} \\
&\qquad - \frac{2(k_1 k_2 - 1)}{k_1(k_1 - 1)}\sum_{u=1}^{k_1}(k_1 - u)\rho_{uv} + \frac{2(k_1 k_2 n_2 - 1)}{k_2 n_2(k_2 n_2 - 1)}\sum_{v=1}^{k_2 n_2}(k_2 n_2 - v)\rho_{0v} \\
&\qquad \left. - \frac{2(k_1 k_2 - 1)}{k_2(k_2 - 1)}\sum_{v=1}^{k_2}(k_2 - v)\rho_{uv}\right].
\end{aligned} \tag{15}
$$

Thus we can evaluate $\sigma^2(X)$ for all types of stratified random sampling.

5c. *Systematic sampling.* In a similar manner to that used for stratified random sampling, we can use (13) to evaluate the variances of systematic sampling. Values of $E(x_i' - \mu)\,(x_j' - \mu)$ for three of the possible methods of sampling are given below. For sy_1sy_1

$$
E(x_i' - \mu)(x_j' - \mu) = \frac{1}{n_1 n_2(n_1 n_2 - 1)}\sum\sum (n_1 - |\, u\,|)(n_2 - |\, v\,|)\rho_{k_1 u, k_2 v} \tag{16}
$$

For sy_1r_0

$$
\begin{aligned}
&E(x_i' - \mu)(x_j' - \mu) = \frac{1}{k_2^2 n_1 n_2(n_1 n_2 - 1)}\sum\sum (n_1 - |\, u\,|) \\
&\qquad \cdot (k_2 n_2 - |\, v\,|)\rho_{k_1 u, v} - \frac{2(k_2 - 1)}{k_2^2 n_2(n_1 n_2 - 1)(k_2 n_2 - 1)}\sum_{v=1}^{k_2 n_2}(k_2 n_2 - v)\rho_{0v}.
\end{aligned} \tag{17}
$$

For sy_0sy_0

$$
\begin{aligned}
&E(x_i' - \mu)(x_j' - \mu) = \frac{1}{k_1 k_2(n_1 n_2 - 1)}\left[\frac{1}{k_1 k_2 n_1 n_2}\sum\sum (k_1 n_1 - |\, u\,|)\right. \\
&\qquad \cdot (k_2 n_2 - |\, v\,|)\rho_{uv} - \frac{1}{k_1 k_2 n_1}\sum\sum (k_1 n_1 - |\, u\,|)(k_2 - |\, v\,|)\rho_{uv} \\
&\qquad - \frac{1}{k_1 k_2 n_2}\sum\sum (k_1 - |\, u\,|)(k_2 n_2 - |\, v\,|)\rho_{uv} + \frac{1}{k_1 k_2}\sum\sum (k_1 - |\, u\,|) \\
&\qquad \left. \cdot (k_2 - |\, v\,|)\rho_{uv} + \frac{k_1}{k_2 n_1}\sum\sum (n_1 - |\, u\,|)(k_2 - |\, v\,|)\rho_{k_1 u, v}\right.
\end{aligned} \tag{18}
$$

$$- \frac{2k_1}{k_2} \sum_{v=1}^{k_2} (k_2 - v)\rho_{0v} + \frac{k_2}{k_1 n_2} \sum\sum (k_1 - |u|)(n_2 - |v|)\rho_{u_1 k_2 v} - \frac{2k_2}{k_1} \sum_{u=1}^{k_1} (k_1 - u)\rho_{u0} \Big].$$

The derivation of (18) may be compared with that of (15).

6. Effect of alignment. We can examine the effect of alignment either by an examination of the values of the variance of different samples, or by the direct use of (13). For random and stratified random sampling, the effect of alignment is to increase the variance of the sample by an amount

$$\Sigma\Sigma\, a_{uv}(\rho_{0v} - \rho_{uv}) + \Sigma\Sigma\, b_{uv}(\rho_{u0} - \rho_{uv}) \qquad \text{where } a_{uv} \geq 0, \; b_{uv} \geq 0.$$

This will be positive for monotonic decreasing correlation functions, and for the majority of functions realised in practice. Thus alignment will usually increase the variance for random and stratified random samples.

For systematic samples, the position is more complicated, but, roughly, the variance is increased by an amount

$$\Sigma\Sigma\, a_{uv}(\rho_{k_1 u, k_2 v} - \bar{\rho}_{k_1 u, k_2 v}),$$

where $a_{uv} \geq 0$ and $\bar{\rho}_{k_1 u, k_2 v}$ is a mean over a rectangle, centre $\rho_{k_1 u, k_2 v}$ for u and v non-zero, and is a mean over a line, length k_1 centre $\rho_{0,\, k_2 v}$ for u zero, (and similarly for v zero). Whether this is positive or negative will depend on the correlation function, and it will have to be investigated for the types of correlation function which are encountered.

7. Limiting forms. For a continuous process, when n_1 and n_2 are large, we may, in the same manner as for linear sampling, obtain integral approximations to the sampling variance, provided that $\Sigma\Sigma\, \rho_{d_1 u, d_2 v}$ converges.

We thus have

$$\sigma^2(r_0 r_0) = \sigma^2(st_0 r_0) \sim \sigma^2/n_1 n_2 , \tag{19}$$

$$\sigma^2(r_1 r_0) \sim \frac{\sigma^2}{n_1 n_2}\Big[1 + \frac{2}{d_2}\int_0^\infty \rho_{0v}\,\delta v\Big]. \tag{20}$$

$$\sigma^2(r_1 r_1) \sim \frac{\sigma^2}{n_1 n_2}\Big[1 + \frac{2}{d_2}\int_0^\infty \rho_{0v}\,\delta v + \frac{2}{d_1}\int_0^\infty \rho_{u0}\,\delta u\Big], \tag{21}$$

$$\sigma^2(st_1 r_0) \sim \frac{\sigma^2}{n_1 n_2}\Big[1 - \frac{1}{d_1^2 d_2}\int_{-\infty}^{\infty}\int_{-d_1}^{d_1} (d_1 - |u|)\rho_{uv}\,\delta u + \frac{2}{d_2}\int_0^\infty \rho_{0v}\,\delta v\Big], \tag{22}$$

$$\sigma^2(st_0 st_0) \sim \frac{\sigma^2}{n_1 n_2}\Big[1 - \frac{1}{d_1^2 d_2^2}\int_{-d_2}^{d_2}\int_{-d_1}^{d_1} (d_1 - |u|)(d_2 - |v|)\rho_{uv}\,\delta u\,\delta v\Big], \tag{23}$$

$$\sigma^2(st_1\, st_1) \sim \frac{\sigma^2}{n_1 n_2}\Bigg[1 - \frac{1}{d_1^2 d_2}\int_{-\infty}^{\infty}\int_{-d_1}^{d_1}(d_1 - |u|)\rho_{uv}\,\delta u\,\delta v$$
$$- \frac{1}{d_1 d_2^2}\int_{-d_2}^{d_2}\int_{-\infty}^{\infty}(d_2 - |v|)\rho_{uv}\,\delta u\,\delta v + \frac{1}{d_1^2 d_2^2}\int_{-d_2}^{d_2}\int_{-d_1}^{d_1}(d_1 - |u|)$$
$$\cdot (d_2 - |v|)\rho_{uv}\,\delta u\,\delta v + \frac{2}{d_1}\int_0^{\infty}\rho_{u0}\,\delta u - \frac{2}{d_1^2}\int_0^{d_1}(d_1 - u)\rho_{u0}\,\delta u$$
$$+ \frac{2}{d_2}\int_0^{\infty}\rho_{0v}\,\delta v - \frac{2}{d_2^2}\int_0^{d_2}(d_2 - v)\rho_{0v}\,\delta v\Bigg], \tag{24}$$

$$\sigma^2(sy_1\, r_0) \sim \frac{\sigma^2}{n_1 n_2}\Bigg[1 - \frac{1}{d_1 d_2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\rho_{uv}\,\delta u\,\delta v$$
$$+ \frac{1}{d_2}\sum_{u=-\infty}^{\infty}\int_{-\infty}^{\infty}\rho_{d_1 u, v}\,\delta v - \frac{2}{d_2}\int_0^{\infty}\rho_{0v}\,\delta v\Bigg], \tag{25}$$

$$\sigma^2(sy_1\, sy_1) \sim \frac{\sigma^2}{n_1 n_2}\Bigg[\sum_{u=-\infty}^{\infty}\sum_{v=-\infty}^{\infty}\rho_{d_1 u, d_2 v} - \frac{1}{d_1 d_2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\rho_{uv}\,\delta u\delta v\Bigg], \tag{26}$$

$$\sigma^2(sy_0\, sy_0) \sim \frac{\sigma^2}{n_1 n_2}\Bigg[1 - \frac{1}{d_1^2 d_2}\int_{-\infty}^{\infty}\int_{-d_1}^{d_1}(d_1 - |u|)\rho_{uv}\delta u\delta v$$
$$- \frac{1}{d_1 d_2^2}\int_{-d_2}^{d_2}\int_{-\infty}^{\infty}(d_2 - |v|)\rho_{uv}\,\delta u\delta v$$
$$+ \frac{1}{d_1^2 d_2^2}\int_{-d_2}^{d_2}\int_{-d_1}^{d_1}(d_1 - |u|)(d_2 - |v|)\rho_{uv}\,\delta u\delta v$$
$$+ \frac{1}{d_2^2}\sum_{u=-\infty}^{\infty}\int_{-d_2}^{d_2}(d_2 - |v|)\rho_{d_1 u, v}\,\delta v - \frac{1}{d_2^2}\int_{-d_2}^{d_2}(d_2 - |v|)\rho_{0v}\,\delta v$$
$$+ \frac{1}{d_1^2}\sum_{v=-\infty}^{\infty}\int_{-d_1}^{d_1}(d_1 - |u|)\rho_{u, d_2 v}\,\delta u - \frac{1}{d_1^2}\int_{-d_1}^{d_1}(d_1 - |u|)\rho_{u0}\,\delta u\Bigg]. \tag{27}$$

8. Particular case where $\rho_{uv} = \rho_u\rho_v$. We note that, if $\rho_{uv} = \rho_u\rho_v$[5] most of these forms can be simplified greatly. If we write

$$sy_u = 1 - \frac{2}{d_1}\int_0^{\infty}\rho_u\,\delta u + 2\sum_{u=1}^{\infty}\rho_{d_1 u},$$
$$st_u = 1 - \frac{2}{d_1^2}\int_0^{d_1}(d_1 - u)\rho_u\,\delta u,$$

with similar forms for sy_v and st_v, and, also

$$f_1 = \frac{2}{d_2}\int_0^{\infty}\rho_v\,\delta v, \qquad f_1' = \frac{2}{d_2^2}\int_0^{d_2}(d_2 - v)\rho_v\,\delta v, \qquad f_1'' = 2\sum_{v=1}^{\infty}\rho_{d_2 v},$$
$$f_2 = \frac{2}{d_1}\int_0^{\infty}\rho_u\,\delta u, \qquad f_2' = \frac{2}{d_1^2}\int_0^{d_1}(d_1 - u)\rho_u\,\delta u, \qquad f_2'' = 2\sum_{u=1}^{\infty}\rho_{d_1 u},$$

[5] A sufficient condition for this to be a valid autocorrelation function is that both ρ_u and ρ_v should be autocorrelation functions.

then we have, for example,

$$\sigma^2(r_1 r_0) \sim \frac{\sigma^2}{n_1 n_2}(1 + f_1), \tag{28}$$

$$\sigma^2(r_1 r_1) \sim \frac{\sigma^2}{n_1 n_2}(1 + f_1 + f_2), \tag{29}$$

$$\sigma^2(st_0 st_0) \sim \frac{\sigma^2}{n_1 n_2}(st_u st_v + st_u + st_v), \tag{30}$$

$$\sigma^2(st_1 st_1) \sim \frac{\sigma^2}{n_1 n_2}(st_u st_v + f_1 st_u + f_2 st_v), \tag{31}$$

$$\sigma^2(sy_1 sy_1) \sim \frac{\sigma^2}{n_1 n_2}(sy_u sy_v + f_1 sy_u + f_2 sy_v), \tag{32}$$

$$\sigma^2(sy_0 sy_0) \sim \frac{\sigma^2}{n_1 n_2}(st_u st_v + f_1' sy_u + f_2' sy_v). \tag{33}$$

From these we get

$$\sigma^2(st_1 st_1) - \sigma^2(sy_1 sy_1) \sim \frac{\sigma^2}{n_1 n_2} \cdot [(st_u st_v - sy_u sy_v) + f_1(st_u - sy_u) + f_2(st_v - sy_v)], \tag{34}$$

$$\sigma^2(sy_1 sy_1) - \sigma^2(st_0 st_0) \sim \frac{\sigma^2}{n_1 n_2} \cdot [\{(1 - sy_u)(1 - sy_v) - (1 - st_u)(1 - st_v)\} + f_1'' sy_u + f_2'' sy_v], \tag{35}$$

$$\sigma^2(st_0 st_0) - \sigma^2(sy_0 sy_0) \sim \frac{\sigma^2}{n_1 n_2}[f_1'(st_u - sy_u) + f_2'(st_v - sy_v)]. \tag{36}$$

The forms (34), (35) and (36) enable us to compare the variances of the samples in two dimensions by using the one-dimensional results. For most practical cases, we know that the f's are positive, $st_u \geq sy_u$ and $st_v \geq sy_v$, so that

$$\sigma^2(st_1 st_1) \geq \sigma^2(sy_1 sy_1) \geq \sigma^2(st_0 st_0) \geq \sigma^2(sy_0 sy_0). \tag{37}$$

The values of $\sigma^2(st_0 st_0)/\sigma^2(r_0 r_0)$, $\sigma^2(sy_1 sy_1)/\sigma^2(r_0 r_0)$, $\sigma^2(sy_0 sy_0)/\sigma^2(r_0 r_0)$ and $\sigma^2(st_0 st_0)/\sigma^2(sy_0 sy_0)$ for $\rho_{d_1 u} = \rho_1^{|u|}$ and $\rho_{d_2 v} = \rho_2^{|v|}$ are given in table 3. It is not difficult to show that for a given number of samples, (d_1, d_2 fixed), $\sigma^2(st_0 st_0)$, $\sigma^2(sy_1 sy_1)$ and $\sigma^2(sy_0 sy_0)$ are least when $\rho_1 = \rho_2$. The expressions tabulated have a value of 1 for $\rho_1 = \rho_2 = 0$ and tend to limiting values of 0, 2/3, 0, and 2 respectively as ρ_1 and ρ_2 tend to 1. It is interesting to note that for ρ_1 and ρ_2 differing by more than 0.4 the grid imposed by $sy_1 sy_1$ is less efficient than purely random sampling. The type of function $\rho_{uv} = \rho_u \rho_v$[6] is, however, less likely to be realised

[6] For a town survey, we might find the correlation between two points depending on a within-streets and a between-streets correlation, so that this function could be realised.

TABLE 3

Comparison of the efficiencies of systematic and random sampling for various values of ρ_1 and ρ_2

ρ_2 \ ρ_1	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	1.000 1.000 1.000 1.000	1.000 1.222 1.000 1.000	1.000 1.500 1.000 1.000	1.000 1.857 1.000 1.000	1.000 2.333 1.000 1.000	1.000 3.000 1.000 1.000	1.000 4.000 1.000 1.000	1.000 5.667 1.000 1.000	1.000 9.000 1.000 1.000	1.000 19.000 1.000 1.000	1.000 ∞ 1.000 1.000
0.1		0.720 0.739 0.596 1.21	0.669 0.754 0.534 1.25	0.632 0.827 0.493 1.28	0.601 0.956 0.462 1.30	0.575 1.160 0.437 1.31	0.551 1.488 0.416 1.32	0.529 2.055 0.398 1.33	0.508 3.215 0.382 1.33	0.489 6.734 0.367 1.33	0.471 ∞ 0.354 1.33
0.2			0.609 0.706 0.462 1.32	0.565 0.721 0.416 1.36	0.529 0.788 0.380 1.39	0.497 0.914 0.352 1.41	0.469 1.134 0.328 1.43	0.443 1.532 0.307 1.44	0.419 2.362 0.289 1.45	0.396 4.911 0.272 1.46	0.375 ∞ 0.257 1.46
0.3				0.516 0.689 0.365 1.41	0.476 0.707 0.327 1.45	0.441 0.778 0.297 1.49	0.409 0.924 0.271 1.51	0.380 1.209 0.249 1.53	0.354 1.825 0.229 1.54	0.328 3.751 0.212 1.55	0.305 ∞ 0.196 1.55
0.4					0.432 0.680 0.288 1.50	0.394 0.702 0.256 1.54	0.360 0.787 0.229 1.57	0.329 0.983 0.206 1.60	0.300 1.437 0.185 1.62	0.272 2.900 0.167 1.63	0.247 ∞ 0.151 1.64
0.5						0.354 0.675 0.223 1.59	0.317 0.703 0.195 1.63	0.284 0.821 0.171 1.66	0.253 1.139 0.150 1.68	0.223 2.228 0.132 1.70	0.196 ∞ 0.115 1.71
0.6							0.279 0.671 0.167 1.67	0.243 0.712 0.142 1.71	0.210 0.908 0.121 1.74	0.180 1.679 0.102 1.76	0.151 ∞ 0.085 1.78
0.7								0.206 0.669 0.118 1.75	0.172 0.742 0.096 1.79	0.139 1.226 0.077 1.82	0.109 ∞ 0.059 1.84
0.8									0.136 0.667 0.074 1.84	0.102 0.863 0.055 1.87	0.070 ∞ 0.037 1.89
0.9										0.067 0.667 0.035 1.92	0.034 ∞ 0.018 1.95

in practice than a centrally-symmetric function, which is independent of the choice of axes. For this reason, we consider next this latter type of function.

9. Centrally-symmetric correlation functions. Dedebant and Wehrte [3] have given a necessary and sufficient condition for $\rho(u, v)$ to be a correlation function as

$$\rho(u, v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cos(\omega u - \mu_v)\delta F(\omega, \mu), \tag{38}$$

or alternatively,

$$f(\omega, \mu) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cos(\omega u - \mu_v)\rho(u, v)\,\delta u\,\delta v. \tag{39}$$

For a centrally-symmetric correlation function we can put $u = r\cos\theta, v = r\sin\theta$ then $\rho(u, v) = \rho(r)$ and

$$f(\omega, \mu) = \frac{1}{(2\pi)^2}\int_0^{\infty}\int_0^{2\pi} \cos(r\sqrt{\omega^2 + \mu^2}\cos\theta_1)\rho(r)r\,d\theta_1\,dr,$$

where $\theta_1 = \theta + \tan^{-1}(\mu/\omega)$,

$$= \frac{1}{2\pi}\int_0^{\infty} J_0(r\tau)\rho(r)r\,dr, \quad \text{where} \quad \tau = \sqrt{\omega^2 + \mu^2}.$$

Thus, if $\rho(u, v)$ is centrally-systematic, then so is $f(\omega, \mu)$ and conversely, so that we get

$$f(\tau) = \frac{1}{2\pi}\int_0^{\infty} J_0(r\tau)\rho(r)r\delta r, \tag{40}$$

and

$$\rho(r) = 2\pi\int_0^{\infty} J_0(r\tau)f(\tau)\tau\delta\tau. \tag{41}$$

We can thus find suitable forms for $\rho(r)$ and $f(\tau)$. In this connection the formula $\int_0^{\infty} J_0(yz)e^{-ay}\delta y = 1/(a^2 + z^2)^{1/2}, a \geq 0$, is useful, since we can see that $\frac{\delta^n}{\delta a^n}(e^{-ay}/y)$ and $\frac{\delta^n}{\delta a^n}(a^2 + z^2)^{-1/2}$ are possible functions for $2\pi f(\tau)$ and $\rho(r)$ although our choice must be limited by the stochastic nature of $\rho(r)$ as well as by its convergence. Thus, for example, $a = n = 0$ gives $1/2\pi\tau$ and $1/r$ as spectral and correlation functions, but these will not converge.

In the linear case, the Markoff process $\rho(u) = e^{-au}$ had a spectral function $f(\tau) = 1/\pi(a^2 + \tau^2)$ which is a Cauchy distribution in one dimension. If we take a two-dimensional Cauchy distribution[7] as our spectral function we get $f(\tau) =$

[7] In the same way as the ordinary Cauchy distribution can be considered as a density distribution on a line produced by a point source at a distance a, radiating in all directions, so can a two-dimensional distribution be considered as a density distribution on a plane from a source at distance a.

$a/2\pi(a^2+\tau^2)^{3/2}$ and $\rho(r) = -\dfrac{\delta}{\delta a}(e^{-ar}/r) = e^{-ar}$. Thus it appears that a generalised Cauchy distribution will be the spectral function for a generalised Markoff process.

We can, of course, consider an "elliptical" Markoff process given by[8]

$$\rho(u, v) = \exp -\left[\frac{u^2}{a^2} - \frac{2muv}{ab} + \frac{v^2}{b^2}\right]^{1/2} \tag{42}$$

but, in what follows, to simplify the computation, m will be taken as zero, so that by changing the units in which d_1 and d_2 are measured, we will work with a process $\rho(r) = e^{-ar}$.

TABLE 4

Comparison of observed serial correlations with theoretical values obtained from a centrally-symmetric correlation function

Distance in miles	Rows		Columns		North-east		South-east	
	Observed	Calculated	Observed	Calculated	Observed	Calculated	Observed	Calculated
1	0.332	0.368	0.310	0.368	—	—	—	—
2	—	—	—	—	0.264	0.243	0.264	0.243
2	0.149	0.135	0.090	0.135	—	—	—	—
$2\sqrt{2}$	—	—	—	—	0.050	0.059	0.129	0.059
3	0.009	0.050	−0.029	0.050	—	—	—	—
$3\sqrt{2}$	—	—	—	—	−0.050	0.018	0.070	0.018
4	0.034	0.018	−0.041	0.018	—	—	—	—
$4\sqrt{2}$	—	—	—	—	−0.020	0.004	0.060	0.004

This process does not seem to be far removed from the type of correlation function experienced in agricultural field work.[9] Osborne [4] has mentioned the possible use of $\rho_u = e^{-\lambda u}$. Mahalanobis [5] has calculated correlations for a paddy field of 800 cells; his values are shown in table 4, together with values of the function e^{-r}. Bearing in mind that the standard error of each of Mahalanobis' values is approximately 0.035, the fit is seen to be quite good, although an elliptical process with axes running south-east and north-east would undoubtedly fit the observations better.

[8] In this light, $\rho(r) = e^{-ar}$ will be called the circular Markoff process, while $\rho_{u,v} = \rho_1^{|u|}\rho_2^{|v|}$ and $\rho_{u,v} = \exp\left\{-\left|\dfrac{u}{a} + \dfrac{v}{b}\right|\right\}$ will be known as degenerate Markoff processes of the first and second orders.

[9] This is further supported by the fact that using a function of this kind it is possible to obtain numerically a law in substantial agreement with Fairfield-Smith's law over a wide range of values.

10. The relative efficiencies of systematic and stratified random sampling. Ideally the correlation functions developed in the last section should be used in the expression (19)–(27), but these functions are not capable of easy integration. An alternative approach can be made if we note that

$$\frac{\sigma^2(st_0 st_0) - \sigma^2(sy_0 sy_0)}{\sigma^2(r_0 r_0)} \sim \frac{1}{d_1}\int_{-d_1}^{d_1}\left(1 - \frac{|u|}{d_1}\right)F(u, d_2)\delta u + \frac{1}{d_2}\int_{-d_2}^{d_2}\left(1 - \frac{|v|}{d_2}\right)F(v, d_1)\delta v \tag{43}$$

where

$$F(u, d_2) = \frac{2}{d_2}\left[\int_0^{d_2}\frac{v}{d_2}\rho_{uv}\delta v + \int_{d_2}^{\infty}\rho_{uv}\delta v - d_2\sum_{v=1}^{\infty}\rho_{u,d_2 v}\right],$$

$$F(v, d_1) = \frac{2}{d_1}\left[\int_0^{d_1}\frac{u}{d_1}\rho_{uv}\delta u + \int_{d_1}^{\infty}\rho_{uv}\delta u - d_1\sum_{u=1}^{\infty}\rho_{d_1 u,v}\right].$$

It is seen that $F(u, d_2)$ and $F(v, d_1)$ are extensions of the expressions obtained for $(\sigma_{st}^2 - \sigma_{sy}^2)/\sigma_r^2$ in section 2. Hence, if $F(u, d_2)$ and $F(v, d_1)$ are both positive functions, systematic sampling is more accurate than stratified random sampling. A particular case of this occurs when $\rho_{uv} = \rho_1^u\rho_2^v$. However when $\rho_{uv} = \exp\{-(u^2 + v^2)^{1/2}\}$, $F(u, d_2)$ is not always positive, since, as u increases, ρ_{uv} becomes a convex function of v. This complicates the interpretation of (43) greatly since it appears that as u varies from 0 to d_1, $F(u, d_2)$ varies from $+\infty$ to an unknown value X. This value will be positive if $d_2 >> d_1$ and negative if $d_1 >> d_2$ so that if the sampling is disproportionate in the two directions systematic sampling will be more efficient than stratified random sampling. Furthermore, if $d_1 = d_2 = d$ and $d \to 0$, $F(u, d) \to \infty$ and systematic sampling again appears to be more efficient. Thus in a wide variety of cases this type of systematic sampling i.e. $sy_0 sy_0$ gives a more accurate result than random sampling.

11. Estimation of sampling errors. An examination of formulas (7)–(18) shows that the principles used for the estimation of linear errors can be used in plane sampling. If we consider that each sample can be broken up into independent units each of which is situated in one of s strata, then for q replications we will have $qr - s$ degrees of freedom for error. For example, $r_0 r_0$, $r_0 r_1$, $st_0 r_0$ and $st_0 r_1$ will have $qn_1 n_2 - 1$, $qn_2 - 1$, $qn_1 n_2 - n_1$ and $qn_2 - 1$ degrees of freedom respectively, so that a single sample will contain an unbiased estimate of error, but $st_0 st_0$, $st_0 st_1$, $st_1 st_1$, $sy_0 sy_0$ and $sy_1 sy_1$ will have $n_1 n_2(q - 1)$, $n_2(q - 1)$, $q - 1$ and $q - 1$ degrees of freedom and will require replication to form a valid estimate of error. We can however use the method of splitting our sample into several parts each of which will give a fairly accurate estimate of error. We may, again, consider the possibility of using a set of systematic samples, which are evenly spaced, to estimate the sampling error, and we will see that the exclusion of the ρ's of lower order may lead to appreciable bias unless the correlation between

successive terms of the sample is small, but, as Yates has pointed out, this method will provide an upper limit for our sampling error. These methods of sampling are illustrated by the examples given below.

12. Examples. We shall consider the three methods of estimating the sampling errors of a systematic sample:

(1) using sets of systematic samples randomly placed with respect to each other, i.e. the material to be sampled is broken up into a series of sub-areas or blocks and several systematic samples are taken in each block; the error variance is calculated from the variances of the systematic samples in each block,

(2) using one set of systematic samples randomly placed, i.e. several systematic samples are taken and the area is then broken up into sub-areas or blocks; the error variance is calculated from the variances of the portions of the systematic samples in each block,

(3) using one systematic sample i.e. one systematic sample is taken which is broken into several systematic samples of wider spacing, e.g. four samples at four times the original spacing, the area is then divided into several sub-areas and the error variance is calculated from the variances of the portions of the sub-systematic samples in each block.

These three methods are increasingly accurate in their estimation of the mean, increasingly biased in their estimation of the sampling variance, and decreasingly difficult in their practical application, so that our method of sampling may vary according to the population and according to the use to which the results are to be put. It is, for example, conceivable that subsequent sampling will yield an improved estimate of error so that initially only a rough guide may be required.

a. If we are sampling from a continuous linear population with a large number of observations in each part into which we split our series, methods (1) and (2) will both give accurate estimates of the variance per term

$$\sigma^2\left(1 - \frac{2}{d}\int_0^\infty \rho_u \delta u + 2\sum_{u=1}^{\infty} \rho_{du}\right).$$

Method (3) will, however, estimate σ^2 instead of the correct variance per term, which is

$$\sigma^2\left(1 - \frac{2q}{d}\int_0^\infty \rho_u \delta u + 2\sum_{u=1}^{\infty} \rho_{du/q}\right).$$

Thus the estimates of sampling variance by method (3) will in general be higher than the estimates by methods (1) and (2), although the actual variance will be lower.

b. Kendall [6, 7] has constructed 480 terms of an artificial series $u_{n+2} = 1.1\, u_{n+1} - 0.5\, u_n + \epsilon_{n+2}$ where the ϵ_n are rectangularly distributed from -49 to 49. For this series $\sigma^2 = 2379.81$ and $s^2 = 2535.11$. The series was split in six parts of 80 terms, for each of which $n = 5$, $k = 16$, $q = 4$, so that 18 degrees of freedom were available for error. The results for this sampling configuration are

given in table 5. The values in this table corroborate the conclusions for large samples of continuous populations.

c. A number of uniformity trials were taken and sampled according to the systems st_1st_1 and sy_1sy_1. For sampling according to the system st_1st_1 the error

TABLE 5

Comparison of three methods of estimating the sampling error of systematic samples for an autoregressive scheme

Method	Estimate of sampling variance per term, s^2, based on 18 degrees of freedom	$E(s^2)$	True sampling variance per term
(1)	3228	2170	2170
(2)	1872	2170	2167
(3)	3709	2577	423

TABLE 6

Comparison of efficiencies of different methods of sampling on three uniformity trials

Source	Kalamkar [8]			Wiebe [9]			Wynne Sayes and Karishna, Iyez [10]		
No. In Cochran's [11] Catalogue	72			132			108		
Crop	Potatoes			Wheat			Sugar cane		
No. of Plots	576			1440			960		
Mean	23.262			587.95			270.89		
Variance per term	15.555			10,918.0*			1794.42		
Type of sampling	$st_1\ st_1$	$sy_1\ sy_1$	$sy_1\ sy_1$	$st_1\ st_1$	$sy_1\ sy_1$	$sy_1\ sy_1$	$st_1\ st_1$	$sy_1\ sy_1$	$sy_1\ sy_1$
Proportion sampled	1/6	1/6	1/6	1/9	1/9	1/9	1/8	1/8	1/8
Method of estimating error		(2)	(3)		(2)	(3)		(2)	(3)
No. of partitions	1	4	4	1	4	4	1	5	5
n_1	3	3	6	4	2	4	4	2	4
k_1	2	2	1	3	6	3	2	4	2
n_2	16	2	4	20	5	10	15	3	6
k_2	6	12	6	6	6	3	8	8	4
q	2	4	1	2	4	1	2	4	1
Mean	23.140	23.435	23.323	586.54	598.65	275.29	275.29	266.72	271.27
Estimated variance per term	9.763	2.689	4.889	5151.6	5772.7	7038.5	1320.15	799.29	1269.54
Degrees of freedom of estimated variance	48	12	12	80	12	12	60	15	15

* Based on the original 1500 plots.

was estimated by taking two samples per strata, while, for sampling according to the system sy_1sy_1, the error was estimated by comparing sets of four samples in each part of the series by methods (2) and (3). The results of this sampling are shown in table 6. While the number of trials is small, the trend to be seen in the results agrees very well with the conclusions reached above.

13. Trend in the population. Frequently in taking samples from a population, we are faced with the problem of a trend. This will not greatly affect random and stratified random samples as estimates of the population mean, but the efficiency of systematic samples will be affected to a large extent. If we consider linear sampling, and denote by S_i the sample whose first element is x_i then the set of samples S_i will usually be monotonic with i and the difference between S_1 and S_k will be large (roughly equal to $x_1 - x_k$).

Yates [1] has suggested a method to overcome this difficulty; by letting S_i represent

$$\frac{1}{n-1}\left[\frac{i}{k}x_i + x_{i+k} + \cdots + x_{i+(n-2)k} + \frac{k-i}{k}x_{n+(n-1)k}\right],$$

the difference between systematic samples due to trend is largely removed. It is easily seen that this necessitates a small loss of information, and in particular, for a continuous random population the variance is $(n - \frac{3}{2})\sigma^2/(n-1)^2$ instead of σ^2/n. For plane samples, the corresponding adjusted sample will be

$$\begin{aligned} S_{ij} = \frac{1}{(n_1-1)(n_2-1)}\Bigg[&\frac{ij}{k_1k_2}x_{ij} + \frac{j}{k_2}x_{i+k_1,j} + \cdots + \frac{j(k_1-i)}{k_1k_2}x_{i+(n_1-1)k_1,j} \\ &+ \frac{i}{k_1}x_{i,j+k_2} + x_{i+k_1,j+k_2} + \cdots + \frac{(k_1-i)}{k_1}x_{i+(n_1-1)k_1,j+k_2} \\ &+ \frac{i(k_2-j)}{k_1k_2}x_{i,j+(n_2-1)k_2} + \cdots + \frac{(k_1-i)(k_2-j)}{k_1k_2}x_{i+(n_1-1)k_1,j+(n_2-1)k_2}\Bigg] \end{aligned}$$

with a similar loss of information.

Trend is, however, most likely to be appreciable in large samples, and in this case, the loss of information due to end adjustments is negligible, so that the conclusions reached above will remain unaltered.

The author wishes to thank Dr. F. Yates and Professor M. S. Bartlett for advice in the preparation of this paper.

REFERENCES

[1] W. G. Cochran, "The relative accuracy of systematic and stratified random samples for a certain class of populations," *Annals of Math. Stat.*, Vol. 17 (1946), p. 164.

[2] F. Yates, "A review of recent statistical developements in sampling and sampling surveys," *Roy. Stat. Soc. Jour.*, Vol. 109 (1946), p. 12.

[3] G. Dedebant and P. Wehrte, "Mecanique aliatoir," *Portugaliae Physics*, Vol. 1 (1945).

[4] J. G. Osborne, "Sampling errors of systematic and random surveys of cover-type areas," *Am. Stat. Asso. Jour.*, Vol. 37 (1942), p. 256.

[5] P. C. Mahalanobis, "On large-scale sample surveys," *Roy. Soc. Phil. Trans.*, B. 231 (1944), p. 329.

[6] M. G. Kendall, "On the analysis of oscillatory time-series," *Roy. Stat. Soc. Jour.*, Vol. 108 (1945), p. 93.

[7] M. G. Kendall, *Contributions to the Study of Oscillatory Time-Series*, Nat. Inst. Econ. Soc. Res., 1946.

[8] R. J. KALAMKAR, "Experimental errors and the field-plot technique with potatoes," *Jour. Agr. Sci.*, 1932, p. 373.

[9] G. A. WIEBE, "Variation and correlation in grain-yield among 1500 wheat nursery plots," *Jour. Agr. Res.*, 1935.

[10] WYNNE SAYER AND P. V. KRISHNA IYER, "On some of the factors that influence the error of field experiments with special reference to sugarcane," *Ind. Jour. Agr. Sci.*, 1936, p. 917.

[11] W. G. COCHRAN, "Catalogue of uniformity trial data," *Roy. Stat. Soc. Suppl. Jour.*, Vol. 4 (1937), p. 233.

[12] F. YATES, "Systematic sampling," *Roy. Soc. Phil. Trans.*, Vol. 241 (1948), p. 345.

REPRESENTATION OF PROBABILITY DISTRIBUTIONS BY CHARLIER SERIES*

By R. P. Boas, Jr.

Brown University

Summary. The paper describes some results concerning the representation of a function by linear combinations of the successive differences of the Poisson distribution, not necessarily the partial sums of the type B series of Charlier.

1. Introduction. For various purposes it is often desired to expand a probability distribution $f(x)$ in a series

$$f(x) \sim \sum_{k=0}^{\infty} c_k \theta_k(x), \tag{1}$$

where the $\theta_k(x)$ are a given set of standard functions. Arguments of a heuristic nature led Charlier [4, 5, 6] to suggest that it would be useful to take the $\theta_k(x)$ in (1) to be either the successive derivatives or the successive differences of some fixed function; the two cases are often referred to as type A series and type B series, respectively. Charlier gave formulas for determining the coefficients in the two cases, but the question of whether the formal series represents the given function in any reasonable sense has to be investigated separately for each particular choice of the function generating the series. Only one special case of each type has been much used: for the A-series, $\theta_0(x)$ is the normal density function $(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}x^2}$; for the B-series, $\theta_0(x)$ is the Poisson function $e^{-\lambda}\lambda^x/x!$ (when x is restricted to take only nonnegative integral values). We shall refer only to these special cases when we speak of A- and B-series in this paper.

There are two distinct problems (which have, however, often been confused) connected with the representation of a function $f(x)$ by a series (1); for convenience, we shall refer to them in this paper as the practical problem and the theoretical problem. In the *practical problem*, we have an empirical function $f(x)$, defined only for a finite number of values of x, which we suspect is representable by $c_0\theta_0(x)$ together with a small correction, so that we hope that a few (say three or four) terms of (1) may give a good representation of $f(x)$ in a relatively simple analytical form with a reasonable amount of computational labor. In some cases, and certainly with the classical A- and B-series which we are considering, we could represent, as closely as desired, any $f(x)$ (however irregular) which takes nonzero values at only a finite number of points; but there is no interest in doing this if the process involves finding too many terms of the series. (Neglect of this fact has led to ill-founded statements by mathematicians about the satisfactory nature of the A- or B-series; but see [27, pp. 38–39].)

Thus it would be of interest to know, if possible, under what circumstances a given empirical density can be represented fairly well by a few terms of a series of a given kind. If no simple criterion can be given, it is desirable to have a means

* Address delivered by invitation at the meeting of the Institute at Boulder, Colorado, on September 1, 1949.

of computing coefficients which will make a few terms of (1) give the best possible fit—best possible being defined in a way appropriate for the problem at hand.

In the *theoretical problem*, $f(x)$ is a function defined for all values of x, or at least for all of an infinite set of equally spaced values of x, arising from theoretical considerations which suggest $c_0\theta_0(x)$ as a reasonable first approximation to $f(x)$. For example, the central limit theorem states that under certain conditions the cumulative distribution function of the sum of a large number of independent random variables is approximately normal; then we might expect that this distribution function would be representable by a series (1) with $\theta_0(x)$ the normal distribution function. For such theoretical purposes we should like to have criteria for the representability of a sufficiently general $f(x)$ by a series (1), where representability is of course to be interpreted appropriately, as ordinary convergence, uniform convergence, convergence in mean square, asymptotic representation, etc., according to the requirements of the problem at hand. The larger the class of $f(x)$ for which we can prove a representation theorem, the larger is the possible domain of applicability of the series to theoretical problems.

2. The A-series. This paper is concerned with the B-series, but for comparison we first mention some properties of the A-series. In the case of the classical A-series, we have the attractive fact that the functions $\theta_n(x)$ are orthogonal with weight function $e^{\frac{1}{2}x^2}$, that is,

$$\int_{-\infty}^{\infty} \theta_n(x)\theta_m(x)e^{\frac{1}{2}x^2}\,dx = 0, \qquad m \neq n.$$

In fact, $e^{\frac{1}{2}x^2}\theta_n(x)$ is, except for a numerical factor, the nth Hermite polynomial. This orthogonality property enables one to compute the coefficients in a series (1) with great ease from

$$n!\,c_n = \int_{-\infty}^{\infty} f(x)\theta_n(x)e^{\frac{1}{2}x^2}\,dx, \tag{2}$$

or since $\theta_n(x)e^{\frac{1}{2}x^2}$ is a polynomial, from the moments of $f(x)$. By the classical theory of orthogonal functions, this means that if the c_n are so computed, and we take $N + 1$ terms of the series, we minimize

$$\int_{-\infty}^{\infty} e^{\frac{1}{2}x^2}[f(x) - F_N(x)]^2\,dx \tag{3}$$

for all possible sums

$$F_N(x) = \sum_{n=0}^{N} c_n\theta_n(x). \tag{4}$$

The convergence theory of Hermite series has been thoroughly investigated by mathematicians, so that it would appear that in theoretical problems, in which $f(x)$ is given for all values of x, we are in a position to find out everything about the representation of $f(x)$ by an A-series. Also in problems of practical curve-

fitting, the fact that the closest approximation to $f(x)$ (in the sense (3)) by sums of the form (4) is given by choosing the coefficients according to (2) seems to leave no more to be said.

However, the formal elegance of the A-series seems to be somewhat misleading. Even when a series converges it by no means follows that its Nth partial sum is the best selection of N terms for representing a given function. Even though the partial sums do give the best fit in the sense of (3), it may not be desirable to measure the closeness of approximation by (3); some other measure of approximation may be better suited to the end in view. For example, it is known that the partial sums of Edgeworth's series (see [8]), which is a rearrangement of the A-series, are more satisfactory for some purposes than the partial sums of the A-series with the coefficients determined by (2). More precisely, Edgeworth's series furnishes an asymptotic expansion, with a remainder term whose order of magnitude can be estimated quite precisely, in circumstances where the series of orthogonal functions does not do this. Again, for practical purposes a few terms of the A-series sometimes exhibit undesirable properties (such as negative frequencies). If $f(x)$ is a function defined only for integral values of x, A. Fisher [10] has suggested and applied the idea of minimizing, not (3), but the sum $\sum_{-\infty}^{\infty} |f(x) - F_n(x)|^2$ in order to determine the coefficients of the approximating sums.

3. The B-series. We can now see how the status of the B-series resembles or differs from that of the A-series. Here we deal principally with a function defined for integral values of x; $\theta_0(x) = \theta(x) = e^{-\lambda}\lambda^x/x!$, $\Delta\theta(x) = \theta(x) - \theta(x-1)$, $\Delta^k\theta(x) = \Delta(\Delta^{k-1}\theta(x))$ and $\theta_k(x) = \Delta^k\theta(x)$; $\theta(x)$ is taken to be 0 for negative integral x. We shall refer to this as the *discrete case* of the B-series. The literature of the subject contains a number of rather painful attempts to put the coefficients into usable form, persisting even after the simple formula

$$c_n = (1/n!) \sum_{l=0}^{n} \binom{n}{l} (-1)^l \lambda^{n-l} \mu_l \tag{5}$$

had been obtained, where μ_n is the nth factorial moment,

$$\mu_n = \sum_{k=n}^{\infty} f(k)k!/(k-n)!.$$

Formula (5) can be derived, for example, by using orthogonality properties of the $\theta_r(x)$. We have, in fact, that $\sum_{x=0}^{\infty} \theta_n(x)\theta_m(x)/\theta_0(x)$ is 0 or $n!\,\lambda^{-n}$ according as $n \neq m$ or $n = m$.

The parameter λ in the B-series is at our disposal, and can for example be chosen in such a way as to improve the convergence of the series. For purposes of practical curve-fitting, it has been customary to choose λ equal to the mean of the distribution $f(x)$, a choice which makes the coefficient c_1 of $\Delta\theta$ equal to zero. Charlier also suggested other methods in which c_1 and c_2, or c_1, c_2 and c_3 are zero [7]. Such choices, of course, may reduce the amount of computation needed

to make use of a given number of differences in fitting a curve; aside from this consideration their use seems to depend on the belief that one improves the convergence of a series by adjusting any available parameters so that as many as possible of the initial terms of the series are zero. This belief does not always seem to be confirmed by the facts. (In particular, compare columns 2 and 5 of Table 1, columns 2 and 4 of Table 2, or columns 2 and 4 of Table 3.)

The theoretical problem of what $f(x)$ can be represented by convergent B-series has been studied by several authors [12, 13, 17, 19, 20, 21, 23, 24, 26, 28]; the study by Schmidt [24; see also 25 and 17] gives necessary and sufficient conditions for the representation in the case of a nonnegative $f(x)$, so that, at least in all cases of interest in statistics, the theoretical problem seems to be completely solved. However, one of the purposes of the present paper is to reopen this apparently closed problem.

There is also a *continuous* version of the B-series, which is suggested by the fact that

$$\theta(x) = (2\pi)^{-1} e^{-\lambda} \int_{-\pi}^{\pi} e^{-ixu} \exp(\lambda e^{iu})\, du \tag{6}$$

reduces to the Poisson function $e^{-\lambda}\lambda^x/x!$ for positive integral x (and to 0 for negative integral x). This form of the B-series has not been much used, and its use is subject to suspicion since it has rather peculiar properties. In particular, it cannot represent, in any reasonable sense, a positive function $f(x)$ or one which is too small as $x \to \infty$ [26, 3]; since the functions which present themselves for representation in practice are both positive and small at infinity, the continuous case of the B-series looks unpromising for applications. (See also [27a], 1a.) However, it has been applied [15].

The purpose of this paper is to describe some results on the B-series which have been obtained in a mathematical paper [3], devoted to what we have called the theoretical problem; some contributions to the practical problem will also be given in the present paper. The starting point of this investigation was the question of what happens if one tries to approximate a function, not by the partial sums of the series (1), but by some other combination of the first N functions $\theta_n(x)$, when approximation is taken in the sense of (unweighted) least-squares. This method of approximation seems well adapted to statistical problems, and leads to simpler mathematical work than ordinary point-by-point convergence of the partial sums. The B-series itself gives a least squares approximation with a weight function $1/\theta_0(x)$. We consider here only the classical B-series, when $\theta_0(x) = \theta(x) = e^{-\lambda}\lambda^x/x!$, $\theta_n(x) = \Delta^n\theta_0(x)$; the main results are substantially the same for rather more general cases [3; see also 14, 25]. In addition, here we consider only nonnegative $f(x)$, assumed zero for negative x. Functions which need not be zero for negative x are handled easily by generalizing the B-series to the form [3]

$$f(x) \sim \sum_{-\infty}^{-1} b_n \nabla^n \theta(x) + \sum_{0}^{\infty} a_n \Delta^n \theta(x), \tag{7}$$

where ∇ denotes the advancing difference: $\nabla\theta(x) = \theta(x) - \theta(x+1)$; there seems to be no particular reason (other than a historical one) for preferring one kind of difference to the other. The generalized series (7) might be useful for graduating symmetrical probability distributions, although it does not seem to have been considered in the literature (cf. [1a]).

4. Results: practical problem. Our question takes somewhat different forms in the two cases which we have described as the practical and the theoretical. In the former, we ask what the coefficients $a_n^{(N)}$ should be so that

$$\sum_{x=0}^{\infty} \left| f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) \right|^2 \tag{8}$$

shall be a minimum, where $f(x)$ is an empirically given function and N is a given integer, in general not very large. If N is 0, 1 or 2, that is, if we use 1, 2 or 3 terms, the best choice of the $a_k^{(N)}$ in (8) can be calculated without difficulty.

For $N = 0$, our question is that of finding the best least-squares fit to $f(x)$ by a Poisson distribution $a_0^{(0)} e^{-\lambda}\lambda^x/x!$; the best choice of $a_0^{(0)}$ is then

$$a_0^{(0)} = \left\{ e^{\lambda} \sum_{x=0}^{\infty} f(x)\lambda^x/x! \right\} \Big/ J_0(2i\lambda), \tag{9}$$

where

$$J_0(iy) = 1 + y^2/(2!)^2 + y^4/(4!)^2 + \cdots$$

(J_0 denotes the Bessel function of order 0); on the other hand, the usual formula (5) gives the different coefficient

$$c_0 = \mu_0 = \sum_{x=0}^{\infty} f(x).$$

This, of course, is simpler than (9) to compute, although its use is based on the uncritical assumption that the first term of the series (1) is the best one to take if only one term is to be used. Charlier [7; see also 10, pp. 101–103] suggested a different formula in which one uses, not $\Delta^k\theta(x)$, but $\Delta^k\theta(px+q)$, the parameters p, q, λ being adjusted to make the terms of (1) in $\Delta\theta$, $\Delta^2\theta$, $\Delta^3\theta$ all zero; here $\theta(x)$ is defined when x is not an integer by interpreting $e^{-\lambda}\lambda^x/x!$ as $e^{-\lambda}\lambda^x/\Gamma(x+1)$, and *not* by using formula (6). Table 2 shows that in at least one numerical case (9) gives a better least-squares fit than Charlier's method (and without introducing gamma functions to take care of $\theta(x)$ for fractional x). However, it is not excluded that Charlier's method will give better results in other cases, since with the change of the functions $\theta_n(x)$ the results of this paper cease to apply.

For $N = 1$, we get the best least-squares approximation to $f(x)$ by

$$a_0^{(1)}\theta(x) + a_1^{(1)}\Delta\theta(x)$$

if

$$(10)\qquad a_0^{(1)} = \frac{e^{2\lambda}}{\alpha+\beta}(\Sigma_0+\Sigma_1),$$
$$a_1^{(1)} = \left\{\frac{\beta}{\alpha^2-\beta^2}\Sigma_0 - \frac{\alpha}{\alpha^2-\beta^2}\Sigma_1\right\} e^{2\lambda},$$

where $\Sigma_0 = \sum_{x=0}^{\infty} f(x)\theta(x)$, $\Sigma_1 = \sum_{x=0}^{\infty} f(x)\theta(x-1)$, $\alpha = J_0(2i\lambda)$, $\beta = -iJ_1(2i\lambda)$, the J's again denoting Bessel functions. For $N = 2$, the corresponding formulas involve also $\gamma = -J_2(2i\lambda)$ and $\Sigma_2 = \sum_{x=0}^{\infty} f(x)\theta(x-2)$. They are:

$$(11)\qquad \begin{aligned} e^{-2\lambda}a_0^{(2)} &= \frac{\beta-\alpha}{2\beta^2-\alpha^2-\alpha\gamma}\Sigma_0 + \frac{2\beta-\alpha-\gamma}{2\beta^2-\alpha^2-\alpha\gamma}\Sigma_1 + \frac{\beta-\alpha}{2\beta^2-\alpha^2-\alpha\gamma}\Sigma_2,\\ e^{-2\lambda}a_1^{(2)} &= \frac{\beta\gamma-\alpha\beta+2\beta^2-2\alpha\gamma}{(\alpha-\gamma)(2\beta^2-\alpha^2-\alpha\gamma)}\Sigma_0 + \frac{\alpha+\gamma-2\beta}{2\beta^2-\alpha^2-\alpha\gamma}\Sigma_1\\ &\qquad + \frac{2\alpha^2-2\beta^2+\beta\gamma-\alpha\beta}{2\beta^2-\alpha^2-\alpha\gamma}\Sigma_2,\\ e^{-2\lambda}a_2^{(2)} &= \frac{\alpha\gamma-\beta^2}{(\alpha-\gamma)(2\beta^2-\alpha^2-\alpha\gamma)}\Sigma_0 + \frac{\beta}{2\beta^2-\alpha^2-\alpha\gamma}\Sigma_1\\ &\qquad + \frac{\beta^2-\alpha^2}{(\alpha-\gamma)(2\beta^2-\alpha^2-\alpha\gamma)}\Sigma_2. \end{aligned}$$

The functions $i^n J_n(iy)$ are real for real y, and extensive tables are available [32].

Some numerical examples showing the comparison between graduation by these formulas and by the corresponding number of terms of the B-series are given in Tables 1–3. It will be noticed that (as the theory indicates) one gets a better least-squares fit by formulas (9), (10) or (11) than by a corresponding number of terms of the B-series using the coefficients (5). However, one may not get a better fit if goodness of fit is measured in some other way, e.g. by χ^2. Unfortunately the coefficients calculated by this method increase rapidly in complexity as the number of terms increases, and even the coefficients for $N = 3$ would involve very heavy algebra. Since numerical examples [2] indicate that it is often necessary to go to terms in $\Delta^4\theta$ for a satisfactory fit, it might be worth while to calculate the next few coefficients.

5. Results: theoretical problem. In the case of a theoretical distribution we ask how coefficients should be determined so that

$$(12)\qquad \sum_{x=0}^{\infty}\left| f(x) - \sum_{k=0}^{N} a_k^{(N)}\Delta^k\theta(x)\right|^2$$

will tend to 0 as $N \to \infty$. The convergence to 0 of (12) is a rather strong kind of convergence, since it implies convergence of the approximating sums to $f(x)$, not only for each x, but even uniformly for all x. Of course, the "best" choice of

$a_k^{(N)}$ as above would be expected to give convergence under the weakest hypotheses, but because of the complexity of these coefficients it seems desirable to make (12) only approximately a minimum; this actually makes no difference in the limit, although the approximation is not usually satisfactory for small values of N. To see the connection between the formulas used here and the "classical" formula (5) for the coefficients in (1), we note that (5) can be written

$$a_n = \frac{(-1)^n}{n!}\sum_{k=0}^{\infty} f(k)\,\frac{d^n}{dz^n}\,[z^k\,e^{\lambda(1-z)}]_{z=1}\,; \tag{13}$$

(5) results if we expand the derivative by Leibniz's rule and rearrange the sum. If we expand $e^{-\lambda z}$ in a power series before differentiating in (13), we obtain

$$a_n = (-1)^n\sum_{k=0}^{\infty} f(k)\sum_{l=\max(k,n)}^{\infty}\binom{l}{n}e^{\lambda}(-\lambda)^{l-k}/l! = e^{\lambda}(-1)^n\sum_{l=n}^{\infty}\binom{l}{n}\sum_{k=0}^{l}\frac{(-\lambda)^{l-k}}{(l-k)!}f(k).$$

If now we break this series off at $n = N$ to obtain

$$a_n^{(N)} = e^{\lambda}(-1)^n\sum_{l=n}^{N}\binom{l}{n}\sum_{k=0}^{l}\frac{(-\lambda)^{l-k}}{(l-k)!}\,f(k), \tag{14}$$

we obtain a sequence of approximations to $f(x)$ by sums $\sum_{k=0}^{N} a_k^{(N)}\Delta^k\theta(x)$ which has, in general, much better convergence properties than the partial sums of the B-series with coefficients a_n given by (5). In particular, if $f(x) = 0$ for $x = -1, -2, \cdots$, this sequence of approximations converges to $f(x)$ whenever $\sum_{x=0}^{\infty}|f(x)|^2$ converges; on the other hand, for nonnegative $f(x)$ it is known [24] that the B-series converges if and only if $\lim_{x\to\infty} f(x)2^x x^k = 0$ for $k = 0, 1, 2, \cdots$, a much more restrictive condition. If we demand that the partial sums of the B-series converge in mean square, that is, that (12) tends to zero with $a_k^{(N)}$ independent of N, we have the even more restrictive condition [3] that $\limsup_{x\to\infty}\{f(x)\}^{1/x} \leqq \frac{1}{3}$.

The approximating sums with coefficients (14) have the additional property that they reproduce $f(x)$ exactly for $x = 0, 1, 2, \cdots, N$. One would expect that in general they would then tend to deviate rather widely from $f(x)$ for larger x, and so would not be satisfactory for practical curve-fitting. However, it seems possible that if we fit such a sum not to $f(x)$, but to $f(px + q)$, with suitable integers p and q, thus making the approximation agree with $f(x)$ at a set of values covering the whole range of definition of $f(x)$, it might give a satisfactory fit elsewhere. This possibility has not been investigated; a similar approach using the partial sums of the B-series was suggested by Charlier [7] and Fisher [10].

6. The continuous case of the B-series. In the continuous case we again ask, not when

$$f(x) = \sum_{n=0}^{\infty} a_n\Delta^n\theta(x) \tag{15}$$

with uniform convergence in every finite interval, but when

$$f(x) = \operatorname*{l.i.m.}_{N\to\infty} \sum_{n=0}^{N} a_n^{(N)} \Delta^n \theta(x), \tag{16}$$

which means that

$$\lim_{N\to\infty} \int_{-\infty}^{\infty} \left| f(x) - \sum_{n=0}^{N} a_n^{(N)} \Delta^n \theta(x) \right|^2 dx = 0. \tag{17}$$

For (15) the following negative results are known [26]: if $f(x) \geqq 0$, (15) cannot converge uniformly on every finite interval (unless $f(x) \equiv 0$); the series, if convergent uniformly on every finite interval, cannot converge to $f(x)$ unless the Fourier transform of $f(x)$ vanishes outside $(-\pi, \pi)$, a condition which

TABLE 1

Number of petals on buttercups. $\lambda = .631$

x	1 Observed frequency	2 Calculated 3 terms (formula 5)	3 Calculated 1 term (formula 9)	4 Calculated 2 terms (formula 10)	5 Calculated 3 terms (formula 11)	6 Calculated 3 terms (formula 14)
5	133	134.9	119.9	130.6	132.9	133.0
6	55	51.6	75.6	62.3	55.3	55.0
7	23	22.5	22.5	13.3	22.1	23.0
8	7	9.5	5.0	1.5	8.5	9.1
9	2	2.9	0.8	0.0	2.4	2.6
10	2	0.6	0.1	0.0	0.5	0.5
Total.....	222	222.0	223.9	207.7	221.7	223.2

automatically excludes any $f(x)$ which vanishes for all large $|x|$ or even is too small as $x \to \infty$. Nevertheless, Jørgensen [15] applies the continuous case successfully to practical problems. A possible explanation of this apparent discrepancy is that if the $a_n^{(N)}$ in (16) are properly determined, (16) will be true under fairly general conditions. To be sure, the mean square difference in (17) cannot be made arbitrarily small unless the Fourier transform $g(x)$ of $f(x)$ vanishes outside $(-\pi, \pi)$, but if $|f(x)|^2$ is integrable the difference can be made small if $g(x)$ is itself small. If $g(x)$ does vanish outside $(-\pi, \pi)$, then (16) is true; and in fact the coefficients $a_k^{(N)}$ can be taken the same as in (14), so that the approximating sums depend only on the values of $f(x)$ for integral values of x; these values are known to determine $f(x)$ under our hypotheses on $g(x)$.

7. Discussion of some numerical results. Table 1. Column 2 gives the fit by two terms of the B-series (really three, since the coefficient of $\Delta\theta$ is zero when

formula (5) is used), as calculated by Charlier [7] (that is, using terms through $\Delta^2\theta$). Column 3 gives the best least-squares fit by a single term, i.e., a Poisson distribution, calculated by formula (9); it is clear that this term alone does not represent the observations very well. Column 4 gives the best least-squares fit by terms through $\Delta\theta$. Column 5 gives the best least-squares fit by terms through $\Delta^2\theta$; the improvement over Charlier's fit by the same number of terms is evident by inspection. Column 6 gives, for comparison, the same number of terms calculated by formula (14), which gives an approximation to the best least-squares fit and necessarily reproduces the data exactly for the first three

TABLE 2

Failure of grains of barley. $\lambda = 2.757$

x	1 Observed frequency	2 Calculated 4 terms (Charlier)	3 Calculated 1 term (Formula 9)	4 Calculated 2 terms (Formula 10)	5 Calculated 3 terms (Formula 11)
0	53	63	47.3	49.9	48.4
1	131	139	130.4	134.7	133.4
2	180	174	179.8	181.6	182.3
3	170	151	165.3	163.2	164.3
4	111	111	113.9	110.0	109.8
5	50	60	62.7	59.3	58.1
6	22	32	28.8	26.5	25.2
7	22	14	11.4	10.2	9.3
8	7	6	3.9	3.4	2.9
9	2	2	1.1	1.0	0.8
10	1	0	0.3	0.2	0.2
Total.....	749	752	744.9	740.0	734.7

values of x. The fact that (14) gives good results here is presumably connected with the small size of λ.

Table 2. Column 2 gives the values calculated by Charlier [7] for a fit after the linear transformation $x \to px + q$, with λ, p and q chosen to make the terms in $\Delta\theta$, $\Delta^2\theta$, $\Delta^3\theta$ all zero (the values were read to the nearest integer from Charlier's graph). Column 3 gives the best least-squares single-term fit calculated by formula (9); this is a considerable improvement for $x \leqq 6$, but for the remainder of the table it is rather poor. Column 4 gives the best least-squares fit by two terms; column 5, that by three. The χ^2-test indicates that the graduation is rather poor in all cases.

Table 3. Column 2 gives the classical calculation with terms through $\Delta^2\theta$; this was given by A. Fisher [10] and (more accurately) by Aroian [2]. Columns 3

and 4 give the best least-squares approximations by two and three terms; column 4 is better than column 2, in this sense, as expected. However, column 4 is a poorer fit when tested by χ^2, chiefly because of the poor fit at $x = 0$. It should be noted that two more terms of the B-series give a more satisfactory fit [2].

TABLE 3

α-particles from a bar of polonium. $\lambda = 3.87155$

x	1 Observed frequency	2 Calculated 3 terms (formula 5)	3 Calculated 2 terms (formula 10)	4 Calculated 3 terms (formula 11)
0	57	49.5	51.3	45.2
1	203	201.3	213.3	190.9
2	383	403.4	399.0	393.5
3	525	532.3	524.8	529.8
4	532	520.6	517.2	525.4
5	408	402.6	407.7	409.7
6	273	254.8	267.7	261.9
7	139	137.1	150.6	141.1
8	45	64.0	74.1	65.3
9	27	26.1	32.4	26.3
10	10	9.4	12.8	9.3
11	4	3.0	4.6	2.9
12	0	0.9	1.5	0.8
13	1	0.2	0.5	0.2
14	1	0.0	0.1	0.0
Total.....	2608	2605.2	2657.6	2602.3
		$\chi^2 = 10.2$ $n = 7$	$\chi^2 = 16.2$ $n = 8$	$\chi^2 = 11.4$ $n = 7$

8. Proofs: theoretical problem. We now outline the proofs of the results which we have stated. They depend on the fact that the numbers $\theta(x)$ ($x = 0, \pm 1, \pm 2, \cdots$) (where $\theta(x) = 0$ when x is a negative integer) are the Fourier coefficients of the function $\varphi(u) = e^{-\lambda} \exp(\lambda e^{iu})$, i.e.

$$\theta(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} \varphi(u) e^{-ixu}\, du, \qquad x = 0, \pm 1, \pm 2, \cdots$$

Furthermore,

$$\Delta^k \theta(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} \varphi(u)(1 - e^{iu})^k e^{-ixu}\, du.$$

If we then assume the condition $\sum_{-\infty}^{\infty} |f(x)|^2 < \infty$, with $f(x) = 0$ for $x = -1, -2, \cdots$, the numbers $f(x)$ are the Fourier coefficients of a function $g(x)$ of integrable square, by the Riesz-Fischer theorem from the theory of Fourier series [31, p. 74]:

$$f(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} g(u) e^{-ixu}\, du, \qquad x = 0, \pm 1, \pm 2, \cdots.$$

Thus

$$(18)\quad f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} e^{-ixu} \left[g(u) - \varphi(u) \sum_{k=0}^{N} a_k^{(N)} (1 - e^{iu})^k \right] du,$$

and so the expressions on the left appear as the Fourier coefficients of the expressions in square brackets on the right. By Parseval's theorem for Fourier series [31, p. 76], then, we have

$$(19)\quad \sum_{x=-\infty}^{\infty} \left| f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) \right|^2 = (2\pi)^{-1} \int_{-\pi}^{\pi} \left| g(u) - \varphi(u) \sum_{k=0}^{N} a_k^{(N)} (1 - e^{iu})^k \right|^2 du.$$

Thus we have reduced the problem of minimizing the mean-square difference on the left of (19) to that of minimizing the integral on the right of (19). By rearranging the sum in the integrand, we see that an equivalent problem is to minimize

$$(20)\qquad D = (2\pi)^{-1} \int_{-\pi}^{\pi} \left| g(u) - \varphi(u) \sum_{k=0}^{N} c_k^{(N)} e^{iku} \right|^2 du,$$

where the $c_k^{(N)}$ and $a_k^{(N)}$ are readily expressed in terms of each other; in fact,

$$(21)\qquad a_k^{(N)} = (-1)^k \sum_{l=k}^{N} \binom{l}{k} c_l^{(N)}.$$

Since $|\varphi(u)| = e^{-\lambda + \lambda \cos u} \geqq e^{-2\lambda} > 0$, we can write D in the form

$$D = (2\pi)^{-1} \int_{-\pi}^{\pi} \left| g(u)/\varphi(u) - \sum_{k=0}^{N} c_k^{(N)} e^{iku} \right|^2 |\varphi(u)|^2\, du,$$

so that

$$(22)\qquad \int_{-\pi}^{\pi} \left| g(u)/\varphi(u) - \sum_{k=0}^{N} c_k^{(N)} e^{iku} \right|^2 du \geqq 2\pi D \geqq e^{-4\lambda} \int_{-\pi}^{\pi} \left| g(u)/\varphi(u) - \sum_{k=0}^{N} c_k^{(N)} e^{iku} \right|^2 du,$$

since $e^{-2\lambda} \leqq |\varphi(u)| \leqq 1$. Thus we can make D arbitrarily small if and only if we can make

$$D^* = (2\pi)^{-1} \int_{-\pi}^{\pi} \left| g(u)/\varphi(u) - \sum_{k=0}^{N} c_k^{(N)} e^{iku} \right|^2 du \tag{23}$$

arbitrarily small. Now the Fourier coefficients of $g(u)$ are $f(x)$; those of $1/\varphi(u)$ are $e^{\lambda}(-\lambda)^x/x!$ for $x \geqq 0$, 0 for $x < 0$; by the convolution theorem for Fourier coefficients [31, p. 90] the nth Fourier coefficient of $g(u)/\varphi(u)$ is

$$\sum_{k=0}^{n} f(n-k)e^{\lambda}(-\lambda)^k/k!, \qquad n = 0, 1, 2, \cdots, \tag{24}$$

and zero for $n < 0$. Furthermore, it is well known from the theory of Fourier series that D^* is a minimum if $c_k^{(N)}$ are chosen as the first $N+1$ Fourier coefficients of $g(u)/\varphi(u)$, and that this minimum is arbitrarily small for large enough N if and only if the Fourier coefficients of $g(u)/\varphi(u)$ are zero for negative indices —which is in fact the case. If we then take the values (24) for $c_k^{(N)}$, $k = 0, 1, \cdots, N$, and express $a_k^{(N)}$ in terms of $c_k^{(N)}$ by (21), we arrive at the formula (14).

It will be observed that the minimum D is connected with the minimum D^* by

$$\min D \leqq \max |\varphi(u)| \cdot \min D^* \leqq \min D^* \leqq \frac{\min D}{\min |\varphi(u)|} \leqq e^{2\lambda} \min D,$$

so that all that we can say about the approximation given by (14) with a small N is that it is an upper bound for the best possible mean-square approximation by sums (18), and that the best mean-square approximation is at worst $e^{-2\lambda}$ times it. This means that if D^* is small, so is D; but D^* is not necessarily small even if D is. Hence we cannot in general expect the coefficients (14) to be suitable for practical curve-fitting, since they may increase the mean-square error by a factor of as much as $e^{2\lambda}$; we may, however, expect (14) to be better when λ is small.

Now, as we have already observed,

$$f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x)$$

is the xth Fourier coefficient of

$$g(u) - \varphi(u) \sum_{k=0}^{N} a_k^{(N)} (1 - e^{it})^k ;$$

if we write (18) in the form

$$f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) = \int_{-\pi}^{\pi} e^{-ixt} \left[g(t)/\varphi(t) - \sum_{k=0}^{N} a_k^{(N)} (1 - e^{it})^k \right] \varphi(t)\, dt, \tag{25}$$

and choose the $a_k^{(N)}$ as specified above, the expression in square brackets is $g(t)/\varphi(t)$ minus the first $N+1$ terms of its Fourier series, and so the Fourier

series of $[\cdots]$ involves no e^{ikt} with $k < N + 1$. Since the Fourier series of $\varphi(t)$ involves no e^{ikt} with $k < 0$, the product $\varphi(t)[\cdots]$ also involves no e^{ikt} with $k < N + 1$, and therefore the integral in (25) is zero for $x = 0, 1, 2, \cdots, N$ (since it represents the xth Fourier coefficient of $\varphi(t)[\cdots]$). In other words,

$$f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) = 0, \qquad x = 0, 1, 2, \cdots, N.$$

Furthermore, we can compute $f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x)$ for $x > N$ by the convolution formula from the Fourier series of $\varphi(t)$ and $[\cdots]$; for $n > N$, the nth Fourier coefficient of $[\cdots]$ is just that of $g(t)/\varphi(t)$, given by (24), and that of $\varphi(t)$ is $e^{-\lambda}\lambda^n/n!$, so for $x > N$

$$f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) = \sum_{l=N+1}^{x} \left(\sum_{k=0}^{l} f(l-k) e^{\lambda} (-\lambda)^k / k! \right) \theta(x-l)$$

and in particular

$$f(N+1) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(N+1) = \sum_{k=0}^{N+1} f(N+1-k)(-\lambda)^k / k!.$$

9. Proofs: practical problem. We have so far obtained only an estimate for the minimum of D, by obtaining the minimum of D^*; this estimate is satisfactory for large N and so for theoretical purposes. However, to obtain precisely the best mean-square approximation to $f(x)$ by a small number N of terms of the sum in (18), we have to choose $a_k^{(N)}$ so that

$$\sum_{k=0}^{N} a_k^{(N)} (1 - e^{it})^k \varphi(t)$$

is the first $N + 1$ terms of the expansion of $g(t)$ in terms of the set of functions obtained by replacing $(1 - e^{it})^k \varphi(t)$, $k = 0, 1, 2, \cdots$, by an equivalent orthonormal set. The process for obtaining this orthonormal set is well known; it turns out that the integrals that have to be evaluated are expressible in terms of Bessel functions of imaginary argument; the result is that the first orthonormal functions are

$$\psi_0(t) = (2\pi)^{-\frac{1}{2}} \alpha^{-\frac{1}{2}} \exp(\lambda e^{it}),$$

$$\psi_1(t) = (2\pi)^{-\frac{1}{2}} \frac{\alpha_1 - \alpha_0 e^{it}}{[\alpha_0(\alpha_0^2 - \alpha_1^2)]^{\frac{1}{2}}} \exp(\lambda e^{it}),$$

$$\psi_2(t) = (2\pi)^{-\frac{1}{2}} \frac{\alpha_1^2 - \alpha_0\alpha_2 - \alpha_1(\alpha_0 - \alpha_2)e^{it} - (\alpha_1^2 - \alpha_0^2)e^{2it}}{[(\alpha_1^2 - \alpha_0^2)(\alpha_0 - \alpha_2)(2\alpha_1^2 - \alpha_0^2 - \alpha_0\alpha_2)]^{\frac{1}{2}}} \exp(\lambda e^{it}),$$

where $\alpha_0 = J_0(2i\lambda)$, $\alpha_1 = -iJ_1(2i\lambda)$, $\alpha_2 = -J_2(2i\lambda)$. It is then a simple matter, first to express ψ_0, ψ_1, ψ_2 in terms of $\varphi(t)$, $\varphi(t)(1 - e^{it})$, $\varphi(t)(1 - e^{it})^2$, and then to determine $a_0^{(0)}$; $a_0^{(1)}$, $a_1^{(1)}$; and $a_0^{(2)}$, $a_1^{(2)}$, $a_2^{(2)}$. For example, the best two-term

approximation for $g(u)$ in terms of $\psi_0(u)$, $\psi_1(u)$ is

$$g(u) \backsim \psi_0(u) \int_{-\pi}^{\pi} g(u)\bar{\psi}_0(u)\, du + \psi_1(u) \int_{-\pi}^{\pi} g(u)\bar{\psi}_1(u)\, du,$$

and the integrals $\int_{-\pi}^{\pi} g(u)\bar{\psi}_k(u)\, du$ are combinations of terms of the form

$$(2\pi)^{-1} \int_{-\pi}^{\pi} g(u)e^{iku}\, \varphi(u)\, du;$$

these in turn are Fourier coefficients of $g(u)\varphi(u)$ and so are expressible, by the Parseval formula, as products of the Fourier coefficients of $g(u)$ (namely, $f(n)$) and of $\varphi(u)$ (namely, $\theta(n)$). We omit the algebraic work; the results are given in formulas (9), (10), (11).

10. Proofs: continuous case. In the continuous case of our approximation problem we assume that $|f(x)|^2$ is integrable on $(-\infty, \infty)$ and look for coefficients $a_k^{(N)}$ that will minimize

$$D = \int_{-\pi}^{\pi} \left| f(x) - \sum_{k=0}^{N} a_k^{(N)} \Delta^k \theta(x) \right|^2 dx,$$

where

$$\theta(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} \varphi(u)e^{-ixu}\, du,$$

$$\Delta^k \theta(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} \varphi(u)e^{-ixu}(1 - e^{iu})^k\, du.$$

Let $f(x)$ be the Fourier transform of $g(u)$; we can regard $\theta(x)$ as the Fourier transform of $\varphi(u)$, $\varphi(u)$ being defined as zero outside $(-\pi, \pi)$. Then by Parseval's theorem for Fourier transforms we have

$$2\pi D = \int_{|t|>\pi} |g(t)|^2\, dt + \int_{-\pi}^{\pi} \left| g(t) - \varphi(t) \sum_{k=0}^{N} a_k^{(N)}(1 - e^{it})^k \right|^2 dt.$$

Clearly, then, D cannot be made arbitrarily small unless $g(t) = 0$ almost everywhere outside $(-\pi, \pi)$; and if this condition is satisfied, D reduces to the same form which it had in the discrete case—see (19). Thus the problem of mean-square approximation in the continuous case reduces, if it can be solved at all, to the corresponding problem in the discrete case.

11. Representation by a series. We consider the representation of a given $f(x)$ by the B-series with the classical coefficients (5), but with mean-square convergence of the series. Here we assume that $f(x) \geqq 0$, $f(x) = 0$ for $x = -1, -2, \cdots$, and $\sum_{x=0}^{\infty} [f(x)]^2 < \infty$, ask whether we can have

$$\lim_{n\to\infty} \sum_{x=-\infty}^{\infty} \left| f(x) - \sum_{k=0}^{n} a_k \Delta^k \theta(x) \right|^2 = 0, \tag{26}$$

where here the a_k do not depend on n (but are not, in principle, required to have the form (5)). From our previous discussion this is equivalent to

$$\lim_{n\to\infty} \int_{-\pi}^{\pi} \left| g(t) - \varphi(t) \sum_{k=0}^{n} a_k(1 - e^{it})^k \right|^2 dt = 0,$$

and this implies that

$$\lim_{n\to\infty} |a_n|^2 \int_{-\pi}^{\pi} |\varphi(t)|^2 |1 - e^{it}|^{2n} dt = 0.$$

From this it follows easily that

$$\sum_{n=0}^{\infty} a_n(1 - e^{it})^n$$

converges for $|t| < \pi$, or in other words that

$$H(z) = \sum_{n=0}^{\infty} a_n(1 - z)^n$$

converges on $|z| = 1$ except perhaps for $z = -1$, and hence converges in $|1 - z| < 2$. By analytic continuation it is easy to identify $H(z)$ with $F(z)\Phi(z)$, where for $|z| < 1$,

$$F(z) = \sum_{n=0}^{\infty} f(n)z^n, \qquad \Phi(z) = \sum_{n=0}^{\infty} \theta(n)z^n = e^{\lambda(1-z)}.$$

Since $1/\Phi(z)$ has no singular points, $F(z)$ is analytic in $|1 - z| < 2$ and hence in particular in $0 \leqq x < 3$; since $F(z)$ is a power series with nonnegative coefficients, it has a singular point at the positive real point on its circle of convergence [30, p. 214], and so it must be analytic at least in $|z| < 3$. This gives the restriction $\limsup_{n\to\infty} |f(n)|^{1/n} \leqq \frac{1}{3}$. Nevertheless, as we know, $f(x)$ is represented in mean-square by a sequence of sums of terms $a_k^{(N)}\Delta^k\theta(x)$ even if we assume only that $\Sigma |f(n)|^2$ converges.

In the continuous case, if $f(x) \geqq 0$ and we have

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \left| f(x) - \sum_{k=0}^{n} a_k\Delta^k\theta(x) \right|^2 dx = 0, \tag{27}$$

we must have $g(x) = 0$ almost everywhere outside $(-\pi, \pi)$ and then, as we saw previously, (26) holds also. Now since $f(x) \geqq 0$, $g(t)$ has derivatives of all orders if it has derivatives of all orders at $t = 0$ [29, p. 90] and it is easily seen from this that $g(t)$ is analytic for all real t if it is analytic at $t = 0$. Now on the one hand, unless $f(x) \equiv 0$, $g(t)$ cannot be analytic for all real t if (as we are supposing) $g(t)$ vanishes outside $(-\pi, \pi)$. On the other hand, $H(e^{it}) = g(t)/\varphi(t)$ for real values of t close to 0 and so, if t is regarded as a complex variable, for complex values of t near 0. Since $1/\varphi(t)$ is analytic everywhere, $g(t)$ is analytic at $t = 0$. From this contradiction we infer that a nonnegative $f(x)$ can never be represented in the form (27), although it may perfectly well be represented by

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \left| f(x) - \sum_{k=0}^{n} a_k^{(n)}\Delta^k\theta(x) \right|^2 dx = 0.$$

REFERENCES

I. Charlier series

[1] A. C. AITKEN, *Statistical Mathematics*, Oliver and Boyd, London, 1939, pp. 58–59, 66–67, 76–79.

[1a] W. ANDERSSON, "Short notes on Charlier's method for expansion of frequency functions in series," *Skandinavisk Aktuar. tids.*, Vol. 27 (1944), pp. 16–31.

[2] L. A. AROIAN, "The type B Gram-Charlier series," *Annals of Math. Stat.*, Vol. 8 (1937), pp. 183–192.

[3] R. P. BOAS, JR., "The Charlier B-series," *Am. Math. Soc. Trans.* (to appear).

[4] C. V. L. CHARLIER, "Über das Fehlergesetz," *Ark. Mat. Astr. Fys.*, Vol. 2, no. 8 (1905).

[5] C. V. L. CHARLIER, "Die zweite Form des Fehlergesetzes," *ibid.*, no. 15 (1905).

[6] C. V. L. CHARLIER, "Über die Darstellung willkürlicher Funktionen," *ibid.*, no. 20 (1905).

[7] C. V. L. CHARLIER, "Researches into the theory of probability," *Lund Universitets Årsskrift N. F.*, Afd. 2, Vol. 1, no. 5 (1906).

[8] H. CRAMÉR, *Random Variables and Probability Distributions*, Cambridge Tracts in Mathematics and Mathematical Physics, no. 36, Cambridge University Press, 1937.

[9] W. P. ELDERTON, *Frequency Curves and Correlation*, 3d ed., Cambridge University Press, 1938, pp. 131–132.

[10] A. FISHER, *An Elementary Treatise on Frequency Curves and their Application in the Analysis of Death Curves and Life Tables*, Macmillan, New York, 1922.

[11] A. FISHER, *The Mathematical Theory of Probabilities and Its Application to Frequency Curves and Statistical Methods*, Vol. 1, 2d ed., Macmillan, New York, 1922, pp. 271–279.

[12] M. JACOB, "Über die Charlier'sche B-Reihe," *Skandinavisk Aktuar. tids.*, Vol. 15 (1932), pp. 286–291.

[13] C. JORDAN, "Sur la probabilité des épreuves répétées, le théorème de Bernoulli et son inversion," *Bull. Soc. Math. France*, Vol. 54 (1926), pp. 101–137.

[14] N. R. JØRGENSEN, "Note sur la fonction de répartition de type B de M. Charlier," *Ark. Mat. Astr. Fys.*, Vol. 10, no. 15 (1915).

[15] N. R. JØRGENSEN, "Undersøgelser over Frequensflader og Korrelation," *Copenhagen thesis*, 1916.

[16] M. G. KENDALL, *The Advanced Theory of Statistics*, Vol. 1, 3d ed., Griffin, London, 1947, pp. 154–156.

[17] S. KULLBACK, "On the Charlier type B series," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 574–581; Vol. 19 (1948), p. 127.

[18] R. VON MISES, *Vorlesungen aus dem Gebiete der angewandten Mathematik*, Vol. 1, *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik*, Deuticke, Leipzig and Vienna, 1931, pp. 265–269.

[19] N. OBRECHKOFF, *Sur la loi de Poisson, la série de Charlier et les équations linéaires aux différences finies de premier ordre à coefficients constants*, Actual. Sci. Ind., no. 740, Hermann et Cie., Paris (1938), pp. 35–64.

[20] H. POLLACZEK-GEIRINGER, "Die Charlier'sche Entwicklung willkürlicher Verteilungen," *Skandinavisk Aktuar. tids.*, Vol. 11 (1928), pp. 98–111.

[21] H. POLLACZEK-GEIRINGER, "Über die Poissonsche Verteilung und die Entwicklung willkürlicher Verteilungen," *Zeits. f. Angew. Math. und Mech.*, Vol. 8 (1928), pp. 292–309.

[22] H. L. RIETZ, *Mathematical Statistics*, Carus Mathematical Monographs, no. 3, Chicago, 1927, pp. 60–68, 170–177.

[23] E. SCHMIDT, *Über die Charliersche Entwicklung einer "arithmetischen Verteilung" nach den sukzessiven Differenzen der Poissonschen asymptotischen Darstellungsfunktion*

für die Wahrscheinlichkeit seltener Ereignisse, Sitzungsberichte der Preussischen Akademie der Wissenschaften, Phys.-Math. Kl. 1928, p. 148.

[24] E. SCHMIDT, "Über die Charlier-Jordansche Entwicklung einer willkürlichen Funktion nach der Poissonschen Funktion und ihren Ableitung," *Zeits. f. Angew. Math. und Mech.*, Vol. 14 (1933), pp. 139–142.

[25] H. L. SELBERG, "Über die Darstellung willkürlicher Funktionen durch Charliersche Differenzreihen," *Skandinavisk Aktuar. tids.*, Vol. 25 (1942), pp. 228–246.

[26] H. L. SELBERG, "Über die Darstellung der Dichtefunktion einer Verteilung durch eine Charliersche B-Reihe," *Archiv for Mathematik og Naturvidenskab*, Vol. 46 (1943), pp. 127–138.

[27] J. F. STEFFENSEN, *Some Recent Researches in the Theory of Statistics and Actuarial Science*, Cambridge University Press, 1930, pp. 35–48.

[27a] J. F. STEFFENSEN, "Note sur la fonction de type B de M. Charlier," *Svenska Aktuar. Tids.*, Vol. 3 (1916), pp. 226–228.

[28] J. V. USPENSKY, "On Charles Jordan's series for probability," *Annals of Math.* Series 2, Vol. 32 (1931), pp. 306–312.

II. Other references

[29] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[30] E. C. TITCHMARSH, *The Theory of Functions*, Oxford, 1932.

[31] A. ZYGMUND, *Trigonometrical Series*, Warszawa-Lwów, 1935.

[32] *Table of the Bessel Functions $J_0(z)$ and $J_1(z)$ for Complex Arguments*, prepared by the Mathematical Tables Project, National Bureau of Standards, 2d ed., Columbia University Press, New York, 1947.

HEURISTIC APPROACH TO THE KOLMOGOROV-SMIRNOV THEOREMS[1]

By J. L. Doob

University of Illinois

1. Introduction and summary. Asymptotic theorems on the difference between the (empirical) distribution function calculated from a sample and the true distribution function governing the sampling process are well known. Simple proofs of an elementary nature have been obtained for the basic theorems of Komogorov[2] and Smirnov[3] by Feller,[4] but even these proofs conceal to some extent, in their emphasis on elementary methodology, the naturalness of the results (qualitatively at least), and their mutual relations. Feller suggested that the author publish his own approach (which had also been used by Kac), which does not have these disadvantages, although rather deep analysis would be necessary for its rigorous justification. The approach is therefore presented (at one critical point) as heuristic reasoning which leads to results in investigations of this kind, even though the easiest proofs may use entirely different methods.

No calculations are required to obtain the qualitative results, that is the existence of limiting distributions for large samples of various measures of the discrepancy between empirical and true distribution functions. The numerical evaluation of these limiting distributions requires certain results concerning the Brownian movement stochastic process and its relation to other Gaussian processes which will be derived in the Appendix.

2. The problem. Let $x_1, x_2, \cdots$ be mutually independent random variables with a common distribution function $F(\lambda)$,

$$F(\lambda) = Pr\{x_j \leq \lambda\}.$$

In statistical language $x_1, \cdots, x_n$ form a sample of n drawn from the distribution with distribution function $F(\lambda)$. Let $\nu_n(\lambda)$ be the number of these x_j's which are $\leq \lambda$. According to the strong law of large numbers, for each λ

$$\lim_{n\to\infty} \frac{\nu_n(\lambda)}{n} = F(\lambda) \tag{2.1}$$

with probability 1. For fixed n $\nu_n(\lambda)/n$ is itself a distribution function (which depends on the sample values $x_1, \cdots, x_n$) the *empirical distribution function*, and an elaboration of the argument which led to (2.1) shows that (2.1) is true

[1] Research connected with a probability project at Cornell University under an ONR contract.

[2] *Inst. Ital. Atti., Giorn.*, Vol. 4 (1933), pp. 83–91.

[3] *Rec. Math.* (*Matematiceskii Sbornik*), N.S. 6, Vol. 48 (1939), pp. 3–26, *Bull. Math. Univ. Moscou*, Vol. 2 (1939), fasc. 2.

[4] *Annals of Math. Stat.*, Vol. 19 (1948), pp. 177–189.

uniformly in λ, with probability 1; that is if

$$D_n = \underset{-\infty<\lambda<\infty}{\text{L.U.B.}} \left| \frac{\nu_n(\lambda)}{n} - F(\lambda) \right|, \tag{2.2}$$

then D_n is a random variable and

$$\lim_{n\to\infty} D_n = 0$$

with probability 1.[5] This result would be of limited practical statistical importance except that the distribution of D_n does not depend on the distribution function $F(\lambda)$ if $F(\lambda)$ is continuous. In fact in that case the random variables $F(x_1)$, $F(x_2)$, $\cdots$ are mutually independent and each is uniformly distributed in the interval $(0, 1)$; if $\hat{\nu}_n(\lambda)$ is the number of $F(x_j)$'s $\leq \lambda$, for $j \leq n$,

$$\underset{0\leq\mu\leq 1}{\text{L.U.B.}} \left| \frac{\hat{\nu}_n(\mu)}{n} - \mu \right| = \underset{-\infty<\lambda<\infty}{\text{L.U.B.}} \left| \frac{\nu_n(\lambda)}{n} - F(\lambda) \right|.$$

Thus it is no restriction, replacing x_j by $F(x_j)$ if necessary, in finding the distribution of D_n to assume that $F(\lambda) = \lambda$ for $0 \leq \lambda \leq 1$, and

$$D_n = \underset{0\leq\lambda\leq 1}{\text{L.U.B.}} \left| \frac{\nu_n(\lambda)}{n} - \lambda \right|. \tag{2.2'}$$

The results will hold for D_n defined by (2.2) for any continuous $F(\lambda)$. We shall also consider D_n^+ and D_n^-, defined by

$$\begin{aligned} D_n^+ &= \underset{0\leq\lambda\leq 1}{\text{L.U.B.}} \left[\frac{\nu_n(\lambda)}{n} - \lambda \right], \\ D_n^- &= -\underset{0\leq\lambda\leq 1}{\text{G.L.B.}} \left[\frac{\nu_n(\lambda)}{n} - \lambda \right], \end{aligned} \tag{2.3}$$

and again the results will hold (with the obvious definitions of D_n^+ and D_n^- in the general case) for every continous $F(\lambda)$.

The problem is to find the limiting distributions of (properly normalized) D_n, D_n^+, D_n^- when $n \to \infty$.

3. Derivation of the Kolmogorov and Smirnov theorems. Define

$$x_n(t) = n^{\frac{1}{2}} \left(\frac{\nu_n(t)}{n} - t \right), \qquad 0 \leq t \leq 1.$$

Since $\nu_n(0) = 0$ with probability 1 and $\nu_n(t) - \nu_n(s)$ is the number of successes in n independent trials, with probability $t - s$ of success in each trial, $\nu_n(t) - \nu_n(s)$ has expectation $n(t-s)$ and variance $n(t-s)[1-(t-s)]$. Hence

$$\begin{aligned} E\{x_n(t)\} &= 0, && 0 \leq t \leq 1; \\ E\{[x_n(t) - x_n(s)]^2\} &= (t-s)[1-(t-s)], && 0 \leq s \leq t \leq 1. \end{aligned} \tag{3.1}$$

[5] Cf. M. Fréchet, *Généralités sur les probabilités. Variables aléatoires*, Paris, 1937, pp. 260–261.

Now let $\{x(t)\}$ be a one parameter family of random variables, $0 \leq t \leq 1$ with the following properties:

(a) for each j if $0 \leq t_1 < \cdots < t_j \leq 1$ the j-variate distribution of the random variables $x(t_1), \cdots, x(t_j)$ is Gaussian;

(b) (3.1) holds, that is

$$E\{x(t)\} = 0, \qquad 0 \leq t \leq 1; \tag{3.1'}$$

$$E\{[x(t) - x(s)]^2\} = (t - s)\,[1 - (t - s)], \qquad 0 \leq s \leq t \leq 1.$$

(c) $Pr\{x(0) = 0\} = 1$.

According to the central limit theorem, the j variate distribution of $x_n(t_1), \cdots, x_n(t_j)$ is asymptotically that of $x(t_1), \cdots, x(t_j)$; in fact the normalizing factor $n^{\frac{1}{2}}$ in the definition of $x_n(t)$ and the choice of means and variances in (3.1') were made precisely to bring this about. As far as first and second moments are concerned the $x_n(t)$ and $x(t)$ processes are identical; when $n \to \infty$ the distributions, or at least the j variate ones mentioned, become identical also.

We shall assume, until a contradiction frustrates our devotion to heuristic reasoning, that *in calculating asymptotic $x_n(t)$ process distributions when $n \to \infty$ we may simply replace the $x_n(t)$ processes by the $x(t)$ process.* It is clear that this cannot be done in all possible situations, but let the reader who has never used this sort of reasoning exhibit the first counter example.

The $x(t)$ process has continuous sample functions (cf. Appendix). Define

$$D = \operatorname*{Max}_{0 \leq t \leq 1} |x(t)|,$$

$$D^+ = \operatorname*{Max}_{0 \leq t \leq 1} x(t),$$

$$D^- = -\operatorname*{Min}_{0 \leq t \leq 1} x(t).$$

Then in accordance with our substitution principle $n^{\frac{1}{2}}D_n$, $n^{\frac{1}{2}}D_n^+$, $n^{\frac{1}{2}}D_n^-$ have as n becomes infinite the distributions of D, D^+, D^- respectively. (The latter two are the same because the $-x(t)$ process is stochastically identical with the $x(t)$ process.) Thus these simple qualitative considerations have led to the existence of the limiting distributions derived and evaluated by Kolmogorov, who proved:

THEOREM[6] (Kolmogorov).

$$\lim_{n \to \infty} Pr\{n^{\frac{1}{2}}D_n \geq \lambda\} = 2 \sum_1^\infty (-1)^{m+1} e^{-2m^2\lambda^2}; \tag{3.2}$$

$$\lim_{n \to \infty} Pr\{n^{\frac{1}{2}}D_n^+ \geq \lambda\} = \lim_{n \to \infty} Pr\{nD_n^- \geq \lambda\} = e^{-2\lambda^2}. \tag{3.3}$$

To complete our treatment we shall prove in the Appendix that

$$Pr\{D \geq \lambda\} = 2 \sum_1^\infty (-)^{m+1} e^{-2m^2\lambda^2}; \tag{3.2'}$$

[6] In Feller's paper (*loc. cit.*, p. 178, equation (1.4)) the factor 2 in the exponent was omitted by the printer. The same misprint occurs in Smirnov's table of the values of the series in our (3.2), *Annals of Math. Stat.*, Vol. 19 (1948), pp. 279–281.

(3.3′) $$Pr\{D^+ \geq \lambda\} = Pr\{D^- \geq \lambda\} = e^{-2\lambda^2},$$

so that in fact the above considerations have led not only to the existence but to the evaluation of the asymptotic distributions. (Actually we shall prove somewhat more general results about the $x(t)$ process.)

So much for the Kolmogorov theorems. Smirnov obtained results (also independent of the given continuous distribution function $F(\lambda)$) of a somewhat different nature. Let $x_1^*, x_2^*, \cdots$ be mutually independent random variables with the same individual distributions as the x_j's, that is each distributed uniformly in the interval (0, 1); define $\nu_n^*(\lambda)$ as the number of the first n x_j's which are $\leq \lambda$. Smirnov considered the difference between empirical distribution functions,

$$D_{mn} = \underset{0 \leq \lambda \leq 1}{\text{L.U.B.}} \left| \frac{\nu_m(\lambda)}{m} - \frac{\nu_n^*(\lambda)}{n} \right|,$$

as well as D_{mn}^+ and D_{mn}^- defined in the obvious way. To avoid stressing the obvious we consider only the D_{mn}.

THEOREM (Smirnov). *If* $m, n \to \infty$ *in such a way that* $\frac{m}{n} \to r$, *and if* $N = mn/(m + n)$,

(3.4) $$\lim_{n \to \infty} Pr\{N^{\frac{1}{2}} D_{mn} \geq \lambda\} = 2 \sum_1^\infty (-1)^{m+1} e^{-2m^2\lambda^2}.$$

To derive this result define an $x^*(t)$ process stochastically identical with the $x(t)$ process but independent of it. Then if $x_n^*(t)$ is defined by

$$x_n^*(t) = n^{\frac{1}{2}} \left(\frac{\nu_n^*(t)}{n} - t \right),$$

we identify, in accordance with our heuristic principle the process with variables

$$\{x(t) - r^{\frac{1}{2}} x^*(t)\}$$

with the one with variables

$$\left\{ x_m(t) - \left(\frac{m}{n} \right)^{1/2} x_n^*(t) \right\}.$$

Doing this leads to the fact that the distribution of

$$(N)^{1/2} D_{mn} = \left(\frac{n}{m+n} \right)^{1/2} \underset{0 \leq t \leq 1}{\text{L.U.B.}} \left| x_m(t) - \left(\frac{m}{n} \right)^{1/2} x_n^*(t) \right|$$

converges to that of

$$\left(\frac{1}{1+r} \right)^{1/2} \underset{0 \leq t \leq 1}{\text{Max}} \left| x(t) - (r)^{1/2} x^*(t) \right|.$$

Now the $x(t)$ process and the process with variables

$$\left\{\frac{x(t) - (r)^{1/2}x^*(t)}{(1+r)^{1/2}}\right\}$$

are stochastically identical. Hence we are led to the conclusion that the distribution of $(N)^{1/2}D_{mn}$ converges to that of D, and this is Smirnov's theorem, stated above. (The method we use does not seem applicable to Smirnov's deeper theorems on the number of intersections between empirical and true distribution curves or between pairs of empirical distribution curves.)

APPENDIX

4. The Brownian movement process. Consider any Gaussian stochastic process, with random variables $\{x(t)\}$ where t varies in some interval. That is, we assume that for each t in the interval $x(t)$ is a random variable and that for any $j \geq 1$ if $t_1 < \cdots < t_j$ are in the interval the j variate distribution of $x(t_1), \cdots, x(t_j)$ is Gaussian. In the following we shall always assume that $E\{x(t)\} \equiv 0$. Then the process is determined stochastically by the covariance function

$$r(s, t) = E\{x(s)x(t)\}.$$

In particular, if the range of parameter is the interval $[0, \infty)$ and if

$$r(s, t) = \sigma^2 \operatorname{Min}(s, t), \qquad 0 \leq s, t < \infty,$$

the process is called the Brownian movement process, or sometimes the Wiener process; σ is a positive constant. When considering this process we shall write $\zeta(t)$ instead of $x(t)$. For the $\zeta(t)$ process

$$Pr\{\zeta(0) = 0\} = 1,$$
$$E\{[\zeta(t) - \zeta(s)]^2\} = \sigma^2 \mid t - s \mid,$$

and if $0 \leq s_1 < t_1 < s_2 < t_2$ the increments $x(t_1) - x(s_1)$ and $x(t_2) - x(s_2)$ are mutually independent. We shall use the following properties of this process, of which the first two are well known.

(a) The sample functions are everywhere continuous with probability 1. In the following we can therefore write as if all the sample curves were continuous.

(b) For fixed s

$$Pr\{\operatorname*{Max}_{0<t\leq T}[\zeta(s+t) - \zeta(s)] \geq \lambda\} = 2Pr\{\zeta(s+T) - \zeta(s) \geq \lambda\}.^7 \tag{4.1}$$

(Note that the use of a general initial value s, rather than 0, has not added to the generality and we drop this affectation below.)

(c) If $a \geq 0$, $b > 0$, $\alpha \geq 0$, $\beta > 0$, then

$$Pr\{\text{L.U.B.}_{0\leq t<\infty}[\zeta(t) - (at + b)] \geq 0\} = e^{-2ab/\sigma^2}, \tag{4.2}$$

[7] Due to Bachelier; cf. the proof by P. Lévy, *Comp. Math.*, Vol. 7 (1939), p. 293. One way to prove (*a*) is to prove (4.1) first, with L.U.B. instead of Max, and then use it to calculate the probabilities relevant to (*a*).

$$(4.3)\quad Pr\{\underset{0\le t<\infty}{\text{L.U.B.}}\,[\zeta(t)-(at+b)]\ge 0 \quad or \quad \text{G.L.B.}\,[\zeta(t)+\alpha t+\beta]\le 0\},$$

$$=\sum_{m=1}^{\infty}\{e^{-2[m^2ab+(m-1)^2\alpha\beta+m(m-1)(a\beta+\alpha b)]}$$

$$+e^{-2[(m-1)^2ab+m^2\alpha\beta+m(m-1)(a\beta+\alpha b)]}$$

$$-e^{-2[m^2(ab+\alpha\beta)+m(m-1)\alpha\beta+m(m+1)ab]}$$

$$-e^{-2[m^2(ab+\alpha\beta)+m(m+1)a\beta+m(m-1)\alpha b]}\};$$

in particular ($\alpha = a, \beta = b$)

$$(4.3')\qquad Pr\left\{\underset{0\le t<\infty}{\text{L.U.B.}}\,\frac{|\zeta(t)|}{at+b}\ge 1\right\}=2\sum_{1}^{\infty}(-1)^{m+1}e^{-2m^2ab}.$$

The probability in (4.2) is the probability that a $\zeta(t)$ sample curve will ever reach the line with slope a and ordinate intercept b; the probability in (4.3) is the probability that a sample curve will ever reach either of the indicated halflines, one above and one below the t axis. Since the right hand sides are continuous functions of a, b, α, β we could write >0 instead of ≥ 0 and <0 instead of ≤ 0 on the left, so that these probabilities are also the probabilities that a sample curve will ever rise above the indicated line or leave the indicated angle.

It will be convenient to describe a line by its slope and ordinate intercept; the line $[u, v]$ is the line with slope u and ordinate intercept v. We shall take $\sigma = 1$ in the proof; this is no essential restriction since $\zeta(t)/\sigma$ is the random variable of a process of the same type whose σ is 1.

To prove (4.2) let $\varphi(a, b)$ be the probability on the left, the probability that a sample curve will reach the line $[a, b]$. If $b = b_1 + b_2$, $b_i > 0$, a sample curve which is to reach $[a, b]$ must first reach $[a, b_1]$ and then move up to meet a line with slope a, b_2 units above the first meeting with $[a, b_1]$. Then

$$\varphi(a, b_1 + b_2) = \varphi(a, b_1)\,\varphi(a, b_2).$$

Now $\varphi(a, b) \ge Pr\{\zeta(1) \ge a + b\} > 0$ and $\varphi(a, b)$ is monotone non-increasing in b, for fixed a. The only solution of the functional equation with these properties is

$$\varphi(a, b) = e^{-\psi(a)b}.$$

Now $\varphi(a, b)$ is the probability of reaching $[0, b]$ at some first time s and then going on to the line $[a, b]$ which from the vantage point of the first common point $(s, \zeta(s))$ is the line $[a, as]$. In other words, using (4.1)

$$e^{-\psi(a)b} = -\int_0^{\infty} e^{-\psi(a)as}\, d_s\, Pr\{\underset{0\le t\le s}{\text{Max}}\,\zeta(t)\ge b\}$$

$$=\int_0^{\infty} e^{-\psi(a)as}\,\frac{be^{-(b^2)/2s}}{(2\pi)^{1/2}s^{3/2}}\,ds$$

$$= \frac{2}{\pi^2}\int_0^\infty \exp\left[-s^2 - \frac{b^2 a\psi(a)}{2s^2}\right] ds$$
$$= e^{-b(2a\psi(a))^{1/2}},$$

from which it follows that $\psi(a) = 2a$, and this yields (4.2).

To prove (4.3) we consider first the following general problem: Let $[u_1, v_1]$, $[u_2, v_2]$, $\cdots$, $u_j \geq 0$, $v_j \geq 0$ be a sequence of lines; let $t = t_1$ be the first value of t, if any, at which a sample curve meets $[u_1, v_1]$; if t_1 is defined for a sample curve let t_2 be the first value of $t > t_1$, if any at which the curve meets $[-u_2, -v_2]$; if t_2 is defined for a sample curve, let t_3 be the first value of $t > t_2$, if any, at which the curve meets $[u_3, v_3]$, and so on. Let π_n be the probability that there is a point t_n, in other words the probability that a sample curve meets the lines $[u_1, v_1]$, $[-u_2, -v_2]$, $\cdots$ $[(-1)^{n+1}u_n, (-1)^{n+1}v_n]$ in at least n successive points. We write

$$\pi_n = \pi_n(u_1, v_1, \cdots, u_n, v_n).$$

In particular, according to (4.2)

$$\pi_1(u_1, v_1) = e^{-2u_1v_1}. \tag{4.4}$$

To evaluate π_n, let Q be the point $(t_{n-1}, \zeta(t_{n-1}))$ on the sample curve, and suppose for definiteness that n is even. Starting at Q, if there is a t_n, the curve must finally reach $[-u_n, -v_n]$, that is it must go to a line of slope $-u_n$, which is $u_{n-1}t_{n-1} + v_{n-1} + u_nt_{n-1} + v_n$ units vertically below its initial position Q when $t = t_{n-1}$. According to (4.2) the probability of doing this is

$$e^{-2u_n(u_{n-1}t_{n-1}+v_{n-1}+u_nt_{n-1}+v_n)}.$$

Now we replace the line $[-u_n, -v_n]$ by a line which depends on t_{n-1} but which leaves this probability unchanged; the new line has slope $-(u_{n-1} + u_n)$ and is

$$h = \frac{u_n}{u_{n-1} + u_n}(u_{n-1}t_{n-1} + v_{n-1} + u_nt_{n-1} + v_n)$$

units below Q when $t = t_{n-1}$. Finally we reflect this new line in the line parallel to the t axis through Q. These two changes do not affect the probability we are discussing because the changes of $\zeta(t)$ after t_{n-1} are independent of the changes before and have symmetric distributions. The final line has slope $u_{n-1} + v_{n-1}$ and is h units above Q when $t = t_{n-1}$; it is the line

$$\left[u_{n-1} + u_n, \quad \frac{u_{n-1}v_{n-1} + u_nv_n + 2u_nv_{n-1}}{u_{n-1} + u_n}\right]$$

which does not depend on t_{n-1}. This line lies above $[u_{n-1}, v_{n-1}]$ in the first quadrant, so that if a sample curve reaches it the curve must also intersect $[u_{n-1}, v_{n-1}]$. We have thus proved that

$$\pi_n(u_1, v_1; \cdots; u_n, v_n)$$
$$= \pi_{n-1}\left(u_1, v_1; \cdots; u_{n-2}, v_{n-2}; u_{n-1} + u_n, \frac{u_{n-1}v_{n-1} + u_nv_n + 2u_nv_{n-1}}{u_{n-1} + u_n}\right). \tag{4.5}$$

The fundamental identity (4.5) makes it possible to reduce the evaluation of π_n to π_1 in $n-1$ steps; π_1 is evaluated in (4.4). Thus successive meetings with n lines have been reduced to a meeting with a single line. As a first example suppose

$$u_1 = \cdots = u_n = u, \qquad v_1 = \cdots = v_n = v.$$

Then we have

$$\begin{aligned}\pi_n(u, v; \cdots; u, v) &= \pi_{n-1}(u, v; \cdots; 2u, 2v) = \cdots \\ &= \pi_1(nu, nv),\end{aligned}$$

so that

$$\pi_n(u, v; \cdots; u, v) = e^{-2n^2uv}. \tag{4.6}$$

More generally suppose

$$u_1 = u_3 = \cdots = a, \qquad v_1 = v_3 = \cdots = b,$$

$$u_2 = u_4 = \cdots = \alpha, \qquad v_2 = v_4 = \cdots = \beta.$$

Then we show that for suitably chosen $C_j^{(n)}$'s we have according as n is even or odd

$$\pi_n(a, b; \cdots; \alpha, \beta) = \pi_1\left[\frac{n}{2}(a+\alpha), \frac{C_1^{(n)}ab + C_2^{(n)}\alpha\beta + C_3^{(n)}a\beta + C_4^{(n)}\alpha b}{\frac{n}{2}(a+\alpha)}\right];$$

$$\begin{aligned}&\pi_n(a, b; \cdots; a, b) \\ &= \pi_1\left[\frac{n+1}{2}a + \frac{n-1}{2}\alpha, \frac{C_1^{(n)}ab + C_2^{(n)}\alpha\beta + C_3^{(n)}a\beta + C_4^{(n)}\alpha b}{\frac{n+1}{2}a + \frac{n+1\alpha}{2}}\right].\end{aligned} \tag{4.7}$$

For $n = 1$ this form is correct with

$$C_1^{(1)} = 1, \qquad C_2^{(1)} = C_3^{(1)} = C_4^{(1)} = 0.$$

If now n is even and if the equations are true for n,

$$\pi_{n+1}(a, b; \cdots; a, b) = \pi_2\left(a, b; \frac{n}{2}(\alpha + a), \frac{C_1^{(n)}\alpha\beta + C_2^{(n)}ab + C_3^{(n)}\alpha b + C_4^{(n)}a\beta}{\frac{n}{2}\alpha + a}\right)$$

$$\equiv \pi_1\left(\frac{n+2}{2}a + \frac{n}{2}\alpha, \frac{ab + C_1^{(n)} + C_2^{(n)}ab + C_3^{(n)}\alpha b + C_4^{(n)}a\beta + n(\alpha+a)b}{\frac{n+2}{2}a + \frac{n}{2}\alpha}\right),$$

and comparing this with (4.7) we find that

$$C_1^{(n+1)} = C_2^{(n)} + n + 1,$$

$$\begin{aligned} C_2^{(n+1)} &= C_1^{(n)}, \\ C_3^{(n+1)} &= C_4^{(n)}, \\ C_4^{(n+1)} &= C_3^{(n)} + n, \end{aligned} \qquad (n \text{ even}).$$

If n is odd we find similarly that

$$\begin{aligned} C_1^{(n+1)} &= C_2^{(n)} + n, \\ C_2^{(n+1)} &= C_1^{(n)}, \\ C_3^{(n+1)} &= C_4^{(n)}, \\ C_4^{(n+1)} &= C_3^{(n)} + n + 1. \end{aligned}$$

The solution of these equations is

n even	*n odd*
$C_1^{(n)} = \frac{n^2}{4}$	$C_1^{(n)} = \frac{(n+1)^2}{4}$
$C_2^{(n)} = \frac{n^2}{4}$	$C_2^{(n)} = \frac{(n-1)^2}{4}$
$C_3^{(n)} = \frac{n(n-2)}{4}$	$C_3^{(n)} = \frac{n^2-1}{4}$
$C_4^{(n)} = \frac{n(n+2)}{4}$	$C_4^{(n)} = \frac{n^2-1}{4}$

Then

$$\pi_n = e^{-\frac{1}{2}[n^2ab + n^2\alpha\beta + n(n-2)a\beta + n(n+2)\alpha b]} \qquad (n \text{ even}),$$

(4.8)

$$\pi_n = e^{-\frac{1}{2}[(n+1)^2ab + (n-1)^2\alpha\beta + (n^2-1)a\beta + (n^2-1)\alpha b]} \qquad (n \text{ odd}).$$

We can now prove (4.3). In fact the left side is equal to

$$\pi_1(a, b) + \pi_1(\alpha, \beta) - \pi_2(a, b; \alpha, \beta) - \pi_2(\alpha, \beta; a, b) + \cdots,$$

which gives (4.3), on substituting (4.8). Only (4.3′), which follows from the simple (4.6), is used in the application to the Kolmogorov-Smirnov theorems.

5. Transformations of Gaussian processes to the Brownian movement process. The $\zeta(t)$ process studied in section 4 is so simple that it is important to be able to reduce others to it by elementary changes of variable. For example if the covariance function of a Gaussian process has the form

$$r(s, t) = u(s)v(t), \qquad s < t, \tag{5.1}$$

for s, t in some interval, and if the ratio

$$\frac{u(t)}{v(t)} = a(t)$$

is continuous and monotone increasing, with inverse function $a_1(t)$. We define

$$\zeta(t) = \frac{u[a_1(t)]}{v[a_1(t)]}.$$

With this definition the ζ process is Gaussian and since if $s < t$

$$E\{\zeta(s)\zeta(t)\} = \frac{u[a_1(s)]v[a_1(t)]}{v[a_1(s)]v[a_1(t)]} = a[a_1(s)] = s = \text{Min}(s, t),$$

the ζ process is the Brownian movement process with $\sigma = 1$. This transformation from the x to the ζ process is effected by a combination of a change of variable in t and the application of a variable scaling factor. (Conversely, if such a transformation is applied to the Brownian movement process it is trivial to verify that the new covariance function will have the form (5.1). The Gaussian processes with covariance functions of this form are easily seen to be the Gaussian Markov processes.)

6. The Gaussian process with $r(s, t) = s(1 - t)$. In section 3 the Kolmogorov-Smirnov theorems were reduced to properties of a Gaussian process with parameter t, $0 \leq t \leq 1$, for which

$$Pr\{x(0) = 0\} = 1;$$

$$E\{x(t)\} = 0;$$

$$E\{[x(t) - x(s)]^2\} = (t - s)[1 - (t - s)], \quad 0 \leq s < t \leq 1.$$

Now these equations imply that

$$E\{x(t)^2\} = t(1 - t), \qquad E\{x(s)^2\} = s(1 - s),$$

and combining the set we find that

$$r(s, t) = E\{x(s)x(t)\} = s(1 - t), \qquad 0 \leq s < t \leq 1.$$

This covariance function has the form studied in section 5, and using the transformation of that section

$$\zeta(t) = (t + 1)x\left(\frac{t}{t + 1}\right), \qquad 0 \leq t < \infty,$$

defines a Brownian movement process (with $\sigma = 1$). Then if D, D^+, D^- are defined as in section 3, we have from (4.3′)

$$Pr\{D \geq \lambda\} = Pr\left\{\text{L. U. B.}_{0 \leq t < \infty} \left| \frac{\zeta(t)}{t + 1} \right| \geq \lambda\right\} = \sum_1^\infty (-1)^{m+1} e^{-2m^2\lambda^2},$$

and from (4.2)

$$Pr\{D^+ \geq \lambda\} = Pr\{D^- \geq \lambda\} = e^{-2\lambda^2}.$$

This proves (3.2′) and (3.3′). Note that we could go beyond these results, because of our detailed knowledge of the $x(t)$ process. For example we can evaluate

$$\lim_{n\to\infty} Pr\{(n)^{\frac{1}{2}}D_n^- \leq \lambda_1\,, \qquad (n)^{\frac{1}{2}}D_n^+ \leq \lambda_2\}.$$

If $\lambda_1 = \lambda_2 = \lambda$ the probability is the probability that $(n)^{1/2}D_n \leq \lambda$ which we have already treated. In general it is, in the limit,

$$Pr\{\operatorname*{Min}_{0\leq t\leq 1} x(t) \geq -\lambda_1\,, \operatorname*{Max}_{0\leq t\leq 1} x(t) \leq \lambda_2\}$$

$$= Pr\left\{\text{G.L.B.}_{0\leq t\leq\infty} \frac{\zeta(t)}{t+1} \geq -\lambda_1 \text{ L.U.B.}_{0\leq t<\infty} \frac{\zeta(t)}{t+1} \leq \lambda_2\right\}$$

$$= 1 - \sum_{m=1}^{\infty} \{e^{-2[m^2\lambda_2^2+(m-1)^2\lambda_1^2+2m(m-1)\lambda_1\lambda_2]} + e^{-2[(m-1)^2\lambda_2^2+m^2\lambda_1^2+2m(m-1)\lambda_1\lambda_2]}$$

$$- e^{-2[m^2(\lambda_1^2+\lambda_2^2)+m(m-1)\lambda_1\lambda_2+m(m+1)\lambda_1\lambda_2]} - e^{-2[m^2(\lambda_1^2+\lambda_2^2)+m(m+1)\lambda_1\lambda_2+m(m-1)\lambda_1\lambda_2]}\}$$

$$= 1 - \sum_{m=1}^{\infty} \{e^{-2[m\lambda_2+(m-1)\lambda_1]^2} + e^{-2[(m-1)\lambda_2+m\lambda_1]^2} - 2e^{-2m^2(\lambda_1+\lambda_2)^2}\}$$

obtained by setting $a = b = \lambda_2$, $\alpha = \beta = \lambda_1$ in (4.3).

PEARSONIAN CORRELATION COEFFICIENTS ASSOCIATED WITH LEAST SQUARES THEORY

By Paul S. Dwyer

University of Michigan

1. Introduction and summary. It is well known that the zero-order correlation between the predicted value of a variable and the observed value of the variable is the multiple correlation. It is also well known that the zero-order correlation between the residuals for two different variables, when the prediction is from a common set of variables, is the partial correlation. These considerations naturally lead to a systematic investigation of all the zero-order correlations involving the various variables associated with least squares theory. Such an investigation is the purpose of this paper.

As a result of this study it appears that other zero-order correlations include the multiple alienation coefficient, the part correlation coefficient, and certain other coefficients which, as far as I am aware, have not been previously defined.

The paper first examines the case of a single predicted variable and then continues with the case in which two or more variables are predicted simultaneously. The paper includes (1) a theoretical development of the different coefficients and the relations between them, (2) the expression of the formulas in determinantal form, (3) a matrix presentation of the material, and (4) an outline of the calculational techniques—with illustrations.

It should be made clear at the start that this paper deals with populations (finite or infinite) and not with samples from those populations. The sampling distribution of each of the new correlation coefficients defined in this paper might well become the subject of a later investigation, but first we need to know what these correlation coefficients are.

2. The case of the single predicted variable. Notation, definitions, and basic properties. We suppose that a population consists of N individuals with values $X_{1j}, X_{2j}, \cdots, X_{kj}, Y_j$ for the variables $X_1, X_2, \cdots, X_k, Y$ and that Y is linearly predicted from the X_i by the formula

$$E = Y - \alpha_0 - \alpha_1 X_1 - \alpha_2 X_2 - \cdots - \alpha_k X_k = Y - \underset{\cdot}{Y} \tag{1}$$

by least squares theory. For the purposes of this paper, we use a concise summation notation, ΣQ, in place of the more formal serial notation $\sum_{i=1}^{N} Q_i$ which is preferable to the frequency notation $\sum_{x=a}^{b} Q_x f_x$ and, in the continuous case, $\int_a^b Q_x f_x dx$ Moreover it is desirable that the scales of X and Y be chosen so as

to facilitate the easy determination of the various formulas. If we let

$$y_j = \frac{Y_j - \bar{Y}}{\sqrt{N}\sigma_y}; \qquad x_{ij} = \frac{X_{ij} - \bar{X}_i}{\sqrt{N}\sigma_{x_i}} \tag{2}$$

we have $\Sigma x_i^2 = \Sigma y^2 = 1$ with the resulting correlating formula

$$\rho_{x_i y} = \frac{\Sigma x_i y}{\sqrt{(\Sigma x_i^2)(\Sigma y^2)}} = \Sigma x_i y \quad \text{and} \quad \rho_{x_i x_j} = \Sigma x_i x_j. \tag{3}$$

The transformations (2) when applied to (1) give

$$e = \frac{E}{\sqrt{N}\sigma_Y} = y - (\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) = y - \underline{y} \tag{4}$$

where the β's are standard regression coefficients and e is defined to be $\frac{E}{\sqrt{N}\sigma_Y}$. It is to be noted that the values of x_i, y, e, and $\underline{y}$ are all dimensionless.

The values we wish to correlate are those of X_i, Y, E, $\underline{Y}$ of (1). The zero-order correlations involving these are the same as for x_i, y, e, $\underline{y}$ of (4).

3. Correlations with a single predicted variable. We wish to minimize Σe^2. Differentiating with respect to β_i and equating to zero we get

$$\Sigma e x_i = 0 \tag{5}$$

from which by multiplication by β_i and summation for i,

$$\Sigma e \underline{y} = 0. \tag{6}$$

It follows that

$$\Sigma e^2 = \Sigma e(y - \underline{y}) = \Sigma e y = \Sigma(y - \underline{y})y = \Sigma y^2 - \Sigma y \underline{y} = 1 - \Sigma y \underline{y} \tag{7}$$
$$= 1 - \Sigma(e + \underline{y})\underline{y} = 1 - \Sigma \underline{y}^2.$$

Using (4) and (7), we get

$$\Sigma e^2 = \frac{\Sigma E^2}{N\sigma_Y^2} = \frac{\sigma_E^2}{\sigma_Y^2} = 1 - \Sigma \underline{y}^2$$

so that

$$\Sigma \underline{y}^2 = \frac{\sigma_Y^2 - \sigma_E^2}{\sigma_Y^2}. \tag{8}$$

This is the conventional definition (from least squares theory) of the multiple correlation coefficient, so

$$\rho^2_{y:x_1 x_2 \cdots x_k} = \rho^2_{y(x)} = \Sigma \underline{y}^2 = \Sigma y \underline{y}. \tag{9}$$

Application of (9) to (7) gives

$$\Sigma e^2 = 1 - \rho^2_{y(x)} = \kappa^2_{y(x)} = \kappa^2_{y:x_1 x_2 \cdots x_k} \tag{10}$$

where $\kappa_{y(x)}$ is the multiple alienation coefficient. We now have $\Sigma x_i^2 = 1$, $\Sigma y^2 = 1$, $\Sigma e^2 = \kappa_{y(x)}^2$, and $\Sigma \underline{y}^2 = \rho_{y(x)}^2$, so that we are able to present formulas involving $x_i, y, e, \underline{y}$. We first form the cross products

$$\Sigma xy = \rho_{xy}, \tag{11}$$

$$\Sigma xe = 0, \tag{12}$$

$$\Sigma x\underline{y} = \Sigma x(y + e) = \Sigma xy = \rho_{xy}, \tag{13}$$

$$\Sigma ye = \Sigma y(y - \underline{y}) = \Sigma y^2 - \Sigma y\underline{y} = 1 - \Sigma \underline{y}^2 = \kappa_{y(x)}^2, \tag{14}$$

$$\Sigma y\underline{y} = \Sigma \underline{y}^2 = \rho_{y(x)}^2, \tag{15}$$

$$\Sigma e\underline{y} = 0. \tag{16}$$

We then have

$$\rho_{x_i x_j} = \frac{\Sigma x_i x_j}{\sqrt{(\Sigma x_i^2)(\Sigma x_j^2)}} = \Sigma x_i x_j. \tag{17}$$

$$\rho_{x_i y} = \frac{\Sigma x_i y}{\sqrt{(\Sigma x_i^2)(\Sigma y^2)}} = \Sigma x_i y, \tag{18}$$

$$\rho_{xe} = \frac{\Sigma xe}{\sqrt{(\Sigma x^2)(\Sigma e^2)}} = 0, \tag{19}$$

$$\rho_{x\underline{y}} = \frac{\Sigma x\underline{y}}{\sqrt{(\Sigma x^2)(\Sigma \underline{y}^2)}} = \frac{\Sigma x\underline{y}}{\rho_{y(x)}} = \frac{\rho_{xy}}{\rho_{y(x)}}. \tag{20}$$

It is interesting to note that this is unity in case $k = 1$ for then $\rho_{x\underline{y}} = \rho_{y(x)}$. Otherwise the absolute value of $\rho_{x\underline{y}}$ is larger than that of ρ_{xy}. For this reason this coefficient might be called the *multiple augmented correlation coefficient.*

$$\rho_{e\underline{y}} = \frac{\Sigma e\underline{y}}{\sqrt{(\Sigma e^2)(\Sigma \underline{y}^2)}} = \frac{\kappa_{y(x)}^2}{\kappa_{y(x)}} = \kappa_{y(x)}. \tag{21}$$

Thus the correlation between y and its residual is the multiple alienation coefficient.

$$\rho_{y\underline{y}} = \frac{\Sigma y\underline{y}}{\sqrt{(\Sigma y^2)(\Sigma \underline{y}^2)}} = \sqrt{\Sigma \underline{y}^2} = \rho_{y(x)}. \tag{22}$$

Thus, as is well known, the zero-order correlation between observed and predicted y is the multiple correlation.

$$\rho_{e\underline{y}} = \frac{\Sigma e\underline{y}}{\sqrt{(\Sigma e^2)(\Sigma \underline{y}^2)}} = 0. \tag{23}$$

4. Notation for the general case. We need to extend the notation and the definitions before examining explicit formulas for the more general case of two (or more) predicted variables. Suppose that Y_i and Y_j are the two variables

predicted from the same X's. Then from (4) we write

$$(24)\qquad \begin{aligned} e_i &= \frac{E_i}{\sqrt{N}\sigma_{Y_i}} = y_i - \beta_{i1}x_1 - \beta_{i2}x_2 - \cdots - \beta_{ik}x_k = y_i - \bar{y}_i \\ e_j &= \frac{E_j}{\sqrt{N}\sigma_{Y_j}} = y_j - \beta_{j1}x_1 - \beta_{j2}x_2 - \cdots - \beta_{jk}x_k = y_j - \bar{y}_j. \end{aligned}$$

We then have the two sets of normal equations

$$(25)\qquad \Sigma e_i x = 0 \qquad \Sigma e_j x = 0$$

so that

$$(26)\qquad \begin{aligned} \Sigma e_i y_i = 0 \qquad \Sigma e_j y_i = 0 \\ \Sigma e_i y_j = 0 \qquad \Sigma e_j y_j = 0. \end{aligned}$$

It follows that

$$(27)\qquad \begin{aligned} \Sigma e_i e_j = \Sigma e_i(y_j - \bar{y}_j) = \Sigma e_i y_j = \Sigma(y_i - \bar{y}_i)y_j = \Sigma y_i y_j - \Sigma \bar{y}_i y_j \\ = \Sigma y_i y_j - \Sigma y_i \bar{y}_j = \Sigma y_i y_j - \Sigma \bar{y}_i \bar{y}_j = \rho_{ij} - \Sigma y_i y_j \end{aligned}$$

if we use the notation that $\rho_{ij} = \rho_{y_i y_j}$.

5. The correlations involving more than one predicted variable. In this case the y's, the e's and the $\bar{y}$'s (as well as the x's) can have more than one variable so that the correlation coefficients we need, in addition to those of section 3, are $\rho_{y_i y_j}$, $\rho_{e_i e_j}$, $\rho_{y_i \bar{y}_j}$, $\rho_{y_i e_j}$, $\rho_{y_j e_i}$, $\rho_{\bar{y}_i \bar{y}_j}$, $\rho_{\bar{y}_i y_j}$, $\rho_{e_i y_j}$, and $\rho_{\bar{y}_i e_j}$. We need now only the summed products

$$(28)\qquad \begin{aligned} &\Sigma y_i y_j = \rho_{y_i y_j} = \rho_{ij}, \\ &\Sigma e_i e_j = \rho_{ij} - \Sigma \bar{y}_i \bar{y}_j \qquad \text{as given in (27),} \end{aligned}$$

$$(29)\qquad \Sigma y_i e_j = \Sigma y_i(y_j - \bar{y}_j) = \Sigma y_i y_j - \Sigma y_i \bar{y}_j = \rho_{ij} - \Sigma \bar{y}_i \bar{y}_j,$$

$$(30)\qquad \Sigma \bar{y}_i y_j = \Sigma \bar{y}_i \bar{y}_j,$$

$$(31)\qquad \Sigma e_i \bar{y}_j = 0.$$

We have then

$$(32)\qquad \rho_{ij} = \frac{\Sigma y_i y_j}{\sqrt{(\Sigma y_i^2)(\Sigma y_j^2)}} = \Sigma y_i y_j,$$

$$(33)\qquad \rho_{e_i e_j} = \frac{\Sigma e_i e_j}{\sqrt{(\Sigma e_i^2)(\Sigma e_j^2)}} = \frac{\rho_{ij} - \Sigma \bar{y}_i \bar{y}_j}{\kappa_{i(x)}\kappa_{j(x)}}.$$

This is the partial correlation coefficient.

$$(34)\qquad \rho_{\bar{y}_i \bar{y}_j} = \frac{\Sigma \bar{y}_i \bar{y}_j}{\sqrt{(\Sigma \bar{y}_i^2)(\Sigma \bar{y}_j^2)}} = \frac{\Sigma \bar{y}_i \bar{y}_j}{\rho_{i(x)}\rho_{j(x)}}.$$

This coefficient appears to be new. Since it is the correlation of predicted values, I suggest that it be called the *predictions correlation coefficient.*

$$\rho_{y_i e_j} = \frac{\Sigma y_i e_j}{\sqrt{(\Sigma y_i^2)(\Sigma e_j^2)}} = \frac{\rho_{ij} - \Sigma y_i y_j}{\kappa_{j(x)}}, \tag{35}$$

$$\rho_{e_i y_j} = \frac{\Sigma e_i y_j}{\sqrt{(\Sigma e_i^2)(\Sigma y_j^2)}} = \frac{\rho_{ij} - \Sigma y_i y_j}{\kappa_{i(x)}}. \tag{36}$$

The correlations given by (35) and (36) have been defined previously and are known as part correlation coefficients [1; 213,497].

$$\rho_{y_i y_j} = \frac{\Sigma y_i y_j}{\sqrt{(\Sigma y_i^2)(\Sigma y_j^2)}} = \frac{\Sigma y_i y_j}{\rho_{j(x)}}, \tag{37}$$

$$\rho_{y_i y_j} = \frac{\Sigma y_i y_j}{\sqrt{(\Sigma y_i^2)(\Sigma y_j^2)}} = \frac{\Sigma y_i y_j}{\rho_{i(x)}}. \tag{38}$$

The correlations of (37) and (38) appear to be new. Each is, in a sense, a generalization of the multiple correlation coefficient since it becomes the multiple correlation coefficient when $i = j$. I suggest that it might be called the *cross multiple correlation coefficient*, since it correlates the actual value of one variable with the predicted value of another.

$$\begin{aligned} \rho_{e_i y_j} &= \frac{\Sigma e_i y_j}{\sqrt{(\Sigma e_i^2)(\Sigma y_j^2)}} = 0, \\ \rho_{y_i e_j} &= \frac{\Sigma y_i e_j}{\sqrt{(\Sigma y_i^2)(\Sigma e_j^2)}} = 0. \end{aligned} \tag{39}$$

A summary of definitions and names of Pearsonian correlation coefficients associated with least squares theory is presented in Table I. No name is proposed when the coefficient is identically zero.

6. Relations between the correlations. Many relations exist between the correlations defined in earlier sections. Some of the more interesting of these are obtained by the elimination of $\Sigma y_i y_j$ from formulas involving this term. Thus from (34), (37), and (38) we get

$$\Sigma y_i y_j = \rho_{y_i y_j}\rho_{i(x)}\rho_{j(x)} = \rho_{y_i y_j}\rho_{j(x)} = \rho_{y_i y_j}\rho_{i(x)},$$

and from (33), (35), and (36) we get

$$\rho_{ij} - \Sigma y_i y_j = \rho_{e_i e_j}\kappa_{i(x)}\kappa_{j(x)} = \rho_{y_i e_j}\kappa_{j(x)} = \rho_{e_i y_j}\kappa_{i(x)}.$$

We then have

$$\left.\begin{aligned} \rho_{ij} &- \rho_{y_i y_j}\,\rho_{i(x)}\,\rho_{j(x)} \\ \rho_{ij} &- \rho_{y_i y_j}\,\rho_{j(x)} \\ \rho_{ij} &- \rho_{y_i y_j}\,\rho_{i(x)} \end{aligned}\right\} = \left\{\begin{aligned} &\rho_{e_i e_j}\,\kappa_{i(x)}\,\kappa_{j(x)} \\ &\rho_{y_i e_j}\,\kappa_{j(x)} \\ &\rho_{e_i y_j}\,\kappa_{i(x)} \end{aligned}\right. \tag{40}$$

where the six members may be equated in all possible ways.

Interesting and simple relations can also be obtained by formation of ratios. Thus

$$\frac{\rho_{e_i e_j}}{\rho_{y_i e_j}} = \frac{1}{\kappa_{i(x)}} \qquad \frac{\rho_{e_i e_j}}{\rho_{e_i y_j}} = \frac{1}{\kappa_{j(x)}} \qquad \text{so } \frac{\rho_{y_i e_j}}{\rho_{e_i y_j}} = \frac{\kappa_{i(x)}}{\kappa_{j(x)}}. \tag{41}$$

TABLE I

Definition	Name
	Single predicted variable
$\rho_{x_i x_j}$	Correlation coefficient of zero order
ρ_{xy}	Correlation coefficient of zero order
$\rho_{xe} = 0$	None
$\rho_{xy} = \frac{\rho_{xy}}{\rho_{yx}}$	*Multiple augmented correlation coefficient
$\rho_{ye} = \kappa_{y(x)}$	Multiple alienation coefficient
$\rho_{yy} = \rho_{y(x)}$	Multiple correlation coefficient
$\rho_{ey} = 0$	None
	Two or more predicted variables
$\rho_{y_i y_j}$	Correlation coefficient of zero order
$\rho_{e_i e_j}$	Partial correlation coefficient
$\rho_{y_i y_j}$	*Predictions correlation coefficient
$\rho_{y_i e_j}$	Part correlation coefficient
$\rho_{y_i y_j}$	*Cross multiple correlation coefficient
$\rho_{e_i y_j}$	None

* Proposed name

Similarly

$$\frac{\rho_{y_i y_j}}{\rho_{y_i y_j}} = \frac{\rho_{i(x)}}{\rho_{j(x)}}. \tag{42}$$

The geometric mean of similar coefficients yields such expressions as

$$\sqrt{\rho_{y_i e_j}\,\rho_{e_i y_j}} = \rho_{e_i e_j}\sqrt{\kappa_{i(x)}\,\kappa_{j(x)}}$$
$$\sqrt{\rho_{y_i y_j}\,\rho_{y_i y_j}} = \rho_{y_i y_j}\sqrt{\rho_{i(x)}\,\rho_{j(x)}}. \tag{43}$$

7. Determinantal formulas. The implicit normal equations (5) become when expanded

$$\begin{aligned} \rho_{11}\beta_1 + \rho_{12}\beta_2 + \cdots + \rho_{1k}\beta_k &= \rho_{1y} \\ \rho_{21}\beta_1 + \rho_{22}\beta_2 + \cdots + \rho_{2k}\beta_k &= \rho_{2y} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ \rho_{k1}\beta_1 + \rho_{k2}\beta_2 + \cdots - \rho_{kk}\beta_k &= \rho_{ky} \end{aligned} \tag{44}$$

while $\Sigma y\underline{y} = \Sigma\underline{y}^2 = \rho^2_{\underline{y}(x)}$ becomes

$$\rho_{y1}\beta_1 + \rho_{y2}\beta_2 + \cdots + \rho_{yk}\beta_k = \rho^2_{\underline{y}(x)}. \tag{45}$$

Let Δ be the determinant of the matrix of the solution of the k x's and y. Let Δ' be the corresponding determinant with ρ_{yy} replaced by $\rho^2_{\underline{y}(x)}$. Let Δ_{yy} be the determinant of the correlation matrix of the k x's. Then $\rho^2_{\underline{y}(x)} = \Sigma\underline{y}^2 = \Sigma y\underline{y}$ can be expressed as a function of Δ and Δ_{yy}. If (44) and (45) are to hold simultaneously, then $\Delta' = 0$. Expanding Δ' in terms of the bottom row, we get

$$\Delta' = 0 = \rho^2_{\underline{y}(x)}\Delta_{yy} + \text{"terms"}. \tag{46}$$

Similarly

$$\Delta = \rho_{yy}\Delta_{yy} + \text{"terms"} \tag{47}$$

where the "terms" of (46) and (47) are identical. It follows by subtraction that $\Delta = (1 - \rho^2_{\underline{y}(x)})\Delta_{yy}$ and hence that

$$\Sigma y\underline{y} = \Sigma\underline{y}^2 = \rho^2_{\underline{y}(x)} = 1 - \frac{\Delta}{\Delta_{yy}}. \tag{48}$$

Then

$$\Sigma e^2 = \Sigma e\underline{y} = \kappa^2_{\underline{y}(x)} = 1 - \Sigma\underline{y}^2 = 1 - \left(1 - \frac{\Delta}{\Delta_{yy}}\right) = \frac{\Delta}{\Delta_{yy}}. \tag{49}$$

Correlation formulas of section 3 then appear as

$$\rho_{x\underline{y}} = \frac{\rho_{xy}}{\sqrt{1 - \dfrac{\Delta}{\Delta_{yy}}}}, \tag{50}$$

$$\rho_{e\underline{y}} = \sqrt{\frac{\Delta}{\Delta_{yy}}}, \tag{51}$$

$$\rho_{y\underline{y}} = \sqrt{1 - \frac{\Delta}{\Delta_{yy}}}. \tag{52}$$

In a similar way the normal equations (25) become two sets of normal equations. The first set is like (44) with β_s replaced by $\beta_{y_i s}$ and ρ_{ys} replaced by $\rho_{y_i s}$. The second set is similar with i replaced by j. It is desired to find

$$\Sigma y_i\underline{y}_j = \Sigma\underline{y}_i\underline{y}_j = \rho_{y_j 1}\beta_1 + \rho_{y_j 2}\beta_2 + \cdots + \rho_{y_j k}\beta_k. \tag{53}$$

Now using (53) with (51) as applied to y_i and using the technique of the first part of this section, we get

$$\Delta_{y_i y_j} = \rho_{y_i y_j}\Delta_{y_i y_i \cdot y_j y_j} + \text{"terms"}, \tag{54}$$

$$0 = \Sigma y_i y_j \Delta_{y_i y_i \cdot y_j y_j} + \text{"terms"}, \tag{55}$$

where Δ is the determinant of the matrix of the correlations of the k x's, y_i and

y_j; $\Delta_{y_i y_j}$ is the determinant obtained by deleting the column involving correlations of y_i and the row involving correlations of y_i; $\Delta_{y_i y_i \cdot y_j y_j}$ is the determinant of the matrix of the k x's; and the "terms" in (54) and (55) are identical. It follows that

$$\Sigma \underline{y}_i \underline{y}_j = \rho_{ij} - \frac{\Delta_{y_j y_i}}{\Delta_{y_i y_i \cdot y_j y_j}} \tag{56}$$

and thence

$$\rho_{ij} - \Sigma \underline{y}_i \underline{y}_j = \frac{\Delta_{y_i y_j}}{\Delta_{y_i y_i \cdot y_j y_j}} . \tag{57}$$

The formulas of section (5) then appear in determinant form as follows

$$\rho_{e_i e_j} = \frac{\frac{\Delta_{ij}}{\Delta_{ii \cdot jj}}}{\sqrt{\left(\frac{\Delta_{ii}}{\Delta_{ii \cdot jj}}\right)\left(\frac{\Delta_{jj}}{\Delta_{jj \cdot ii}}\right)}} = \frac{\Delta_{ij}}{\sqrt{\Delta_{ii}\,\Delta_{jj}}} \tag{58}$$

as is well known.

$$\rho_{y_i y_j} = \frac{\rho_{ij} - \frac{\Delta_{ij}}{\Delta_{ii \cdot jj}}}{\sqrt{\left(1 - \frac{\Delta_{ii}}{\Delta_{ii \cdot jj}}\right)\left(1 - \frac{\Delta_{jj}}{\Delta_{jj \cdot ii}}\right)}} . \tag{59}$$

$$\rho_{y_i e_j} = \frac{\frac{\Delta_{ij}}{\Delta_{ii \cdot jj}}}{\sqrt{\frac{\Delta_{ij}}{\Delta_{ii \cdot jj}}}} . \tag{60}$$

$$\rho_{y_i y_j} = \frac{\rho_{ij} - \frac{\Delta_{ij}}{\Delta_{ii \cdot jj}}}{\sqrt{1 - \frac{\Delta_{ii}}{\Delta_{ii \cdot jj}}}} . \tag{61}$$

Formulas for $\rho_{e_i y_j}$ and $\rho_{y_i y_j}$ are similar to (60) and (61).

Modern methods of calculating determinants (2), (3), (4), (5) are advised if calculations are to be made from those formulas.

8. Matrix formulas. A matrix presentation is very useful in exhibiting the general features of this theory and in developing compact and easy methods of calculation with finite populations. The matrix presentation here is similar to that given by the author in a previous article [6].

Let the normal equations (24) be represented by the matrix equation.

$$E = Y - XB = Y - \underline{Y}. \tag{62}$$

Then the sets of normal equations become

$$X'E = 0 \quad \text{or} \quad X'(Y - XB) = 0$$

so that

$$X'XB = X'Y. \tag{63}$$

Now since $XB = \underline{Y}$, (63) can be written as $X'\underline{Y} = X'Y$ and it can be shown that

$$\underline{Y}'Y = Y'\underline{Y} = \underline{Y}'\underline{Y}. \tag{64}$$

But under the assumptions of section 2, $X'X$ is the matrix of the intercorrelations of the X's, $X'Y$ is the matrix of the intercorrelations of the x's and y's and $Y'Y$ is the matrix of the intercorrelations of the y's. Hence (63) can be written

$$R_{xx}B = R_{xy} \tag{65}$$

so that

$$B = R_{xx}^{-1}R_{xy}. \tag{66}$$

If Y is composed of a single variable, B is a single column matrix (vector) but if Y is composed of m variables, B is an m column matrix. It follows at once that

$$\underline{Y}'Y = \underline{Y}'\underline{Y} = B'X'XB = B'R_{xy}B = R'_{xy}R_{xx}^{-1}R_{xx}R_{xx}^{-1}R_{xy} = R'_{xy}R_{xx}^{-1}R_{xy} \tag{67}$$

and that

$$\begin{aligned} E'E = (Y - XB)'E = Y'E = Y'(Y - XB) = Y'Y - Y'\underline{Y} \\ = Y'Y - \underline{Y}'\underline{Y} = R_{yy} - R'_{xy}R_{xx}^{-1}R_{xy}. \end{aligned} \tag{68}$$

It thus appears that the matrix (67) has diagonal terms $\Sigma\underline{y}^2 = \Sigma y\underline{y}$ which are the squares of the multiple correlation coefficients, and that the non-diagonal terms are $\Sigma\underline{y}_i\underline{y}_j = \Sigma y_i\underline{y}_j$. Similarly the matrix (68) has diagonal terms $\Sigma e^2 = \kappa^2_{y(x)}$ and non-diagonal terms $\Sigma e_ie_j = \Sigma e_iy_j$. It follows that all the correlation coefficients defined above may be calculated from the matrices R_{xx}, R_{xy}, R_{yy}, $\underline{Y}'\underline{Y}$, and $E'E$. The matrix (67) might be called the *multiple correlation matrix* and the matrix (68) the *multiple alienation matrix*.

Conventional results are expressed in terms of the correlation matrices R_{xx}, R_{xy}, and R_{yy}. All the correlation coefficients defined in this paper may be expressed in terms of these matrices and the multiple correlation and alienation matrices.

9. Calculational method of determining the multiple correlation and multiple alienation matrices. Various methods might be used in calculating the multiple correlation and alienation matrices from the correlation matrices. One method utilizes the square root method of solving simultaneous equations, which has

recently been presented in a number of places, [7] [8] together with a device which is similar to that used by Aitken [9] in eliminating the back solution. This method solves the equation (65) by forming the auxiliary

$$S_{xx}B = S_{xx}R_{xx}^{-1}R_{xy} \tag{69}$$

where S_{xx} is a triangular matrix such that

$$R_{xx} - S'_{xx}S_{xx} = 0. \tag{70}$$

TABLE II

General		Illustration					
	R_{yy}					1.000 —	.495 1.000
R_{xx}	R_{xy}	1.000 — — —	.652 1.000 — —	.554 .747 1.000 —	.615 .693 .774 1.000	.313 .280 .182 .166	.650 .803 .804 .812
S_{xx}	$S_{xx}R_{xx}^{-1}R_{xy}$	1.000	.652 .758	.554 .509 .659	.615 .385 .360 .586	.313 .100 .064 .072	.650 .500 .287 .199
	$Y'Y$					.117 —	.221 .794
	$E'E$					.883 —	.274 .206

The right hand side of (69), when premultiplied by its transpose yields

$$(S_{xx}R_{xx}^{-1}R_{xy})'(S_{xx}R_{xx}^{-1}R_{xy}) = R'_{xy}R_{xx}^{-1}S'_{xx}S_{xx}R_{xx}^{-1}R_{xy} = R'_{xy}R_{xx}^{-1}R_{xy} = Y'Y. \tag{71}$$

Speaking less technically it is only necessary to multiply the columns of $S_{xx}R_{xx}^{-1}R_{xy}$ to get $Y'Y$.

A first illustration utilizes the correlations of the Carver anthropometric data [10] for 1000 University of Michigan freshmen. This group may be regarded as constituting a population, or it may be regarded as a random sample of a larger population. For present purposes we regard it as a population. Height (Y_1) and weight (Y_2) are estimated from shoulder girth (X_1) chest girth (X_2), waist girth (X_3), and right thigh girth (X_4). The calculation of $Y'Y$ and $E'E$ from the correlation matrices follow.

As a second illustration I use the correlation between the parts of two forms of the Thorndike Intelligence Examination which Lorge has used in illustration canonical correlation technique [11, 69–74]. The X's are the scores on the three parts of Form A and the Y's are the scores on the three parts of Form B. In this case we designate the results by r's and k's (rather than ρ's and κ's) since the calculation is considered to be for a sample. The calculation of the sample multiple correlation and multiple alienation matrices is presented in Table III.

TABLE III

	Form A			Form B			
	x_1	x_2	x_3	y_1	y_2	y_3	
				1.0000	.8235	.7912	
				—	1.0000	.8315	R_{yy}
				—	—	1.0000	
	1.0000	.7830	.7852	.8986	.7841	.8217	
R_{xx}	—	1.0000	.8393	.7961	.8543	.8254	R_{xy}
	—	—	1.0000	.7683	.8226	.8588	
	1.0000	.7830	.7852	.8986	.7841	.8217	
S_{xx}		.6220	.3609	.1487	.3864	.2926	$S_{xx}R_{xx}^{-1}R_{xy}$
			.5032	.0180	.1341	.2146	
				.8299	.7645	.7858	
				—	.7821	.7861	$\underline{Y}'\underline{Y}$
				—	—	.8069	
				.1701	.0590	.0054	
				—	.2179	.0454	$E'E$
				—	—	.1991	

10. The numerical values of the coefficients. The diagonal entries of the multiple correlation matrix give the values of $\Sigma y_i^2 = \Sigma y_i y_i = \rho_{y(x)}^2$ while the non-diagonal values are $\Sigma y_i y_j = \Sigma y_i y_j$. The diagonal entries of the multiple alienation matrix are $\Sigma e_i^2 = \Sigma e_i y_i = \kappa_{y(x)}^2$ while the non-diagonal entries are $\Sigma e_i e_j = \Sigma e_i y_j = \Sigma y_i e_j$. We are then able to write out any of the correlations easily. Thus from Table II

$$\rho_{1(x)} = \sqrt{\Sigma y_1^2} = \sqrt{.117} = .342,$$

$$\rho_{2(x)} = \sqrt{\Sigma y_2^2} = \sqrt{.794} = .891,$$

$$\kappa_{1(x)} = \sqrt{\Sigma e_1^2} = \sqrt{.883} = .940,$$

$$\kappa_{2(x)} = \sqrt{\Sigma e_2^2} = \sqrt{.206} = .454,$$

$$\rho_{12(x)} = \frac{\Sigma e_1 e_2}{\sqrt{(\Sigma e_1^2)(\Sigma e_2^2)}} = \frac{.274}{\sqrt{(.883)(.206)}} = .643,$$

$$\rho_{y_1 y_2} = \frac{\Sigma y_1 y_2}{\sqrt{(\Sigma y_1^2)(\Sigma y_2^2)}} = \frac{.221}{\sqrt{(.117)(.794)}} = .724,$$

$$\rho_{y_1 e_2} = \frac{\Sigma e_1 e_2}{\sqrt{\Sigma e_2^2}} = \frac{.274}{\sqrt{.206}} = .604,$$

$$\rho_{y_2 e_1} = \frac{\Sigma e_1 e_2}{\sqrt{\Sigma e_1^2}} = \frac{.274}{\sqrt{.883}} = .291,$$

$$\rho_{y_1 y_2} = \frac{\Sigma y_1 y_2}{\sqrt{\Sigma y_2^2}} = \frac{.221}{\sqrt{.794}} = .248,$$

$$\rho_{y_1 y_2} = \frac{\Sigma y_1 y_2}{\sqrt{\Sigma y_1^2}} = \frac{.221}{\sqrt{.117}} = .646.$$

TABLE IVa

General			Illustration		
	$r_{y_1 y_2}$	$r_{y_1 y_3}$	.9110	.9489 .8392 | .8644	.9603 .8626 | .8747
$\rho_{1(x)}$ Σy_1^2	$r_{y_1 y_2}$ | $r_{y_1 y_2}$ $\Sigma y_1 y_2$	$r_{y_1 y_3}$ | $r_{y_1 y_3}$ $\Sigma y_1 y_3$	.8299	.7645	.7858
		$r_{y_2 y_3}$			.9917
	$r_{2(x)}$	$r_{y_2 y_3}$ | $r_{y_2 y_3}$		.8844	.8889 | .8751
	Σy_2^2	$\Sigma y_2 y_3$		.7821	.7861
		$r_{3(x)}$			.8983
		Σy_3^2			.8069

TABLE IVb

General			Illustration		
	$r_{e_1 e_2}$	$r_{e_1 e_3}$	.4124	.3066 .1431 | .1264	.0298 .0131 | .0123
$k_{1(x)}$ Σe_1^2	$r_{e_1 y_2}$ | $r_{y_1 e_2}$ $\Sigma e_1 e_2$	$r_{e_1 y_3}$ | $r_{y_1 e_3}$ $\Sigma e_1 e_3$	.1701	.0590	.0054
		$r_{e_2 e_3}$			.2214
	$k_{2(x)}$	$r_{e_2 y_3}$ | $r_{y_2 e_3}$		.4668	.0973 | .1033
	Σe_2^2	$\Sigma e_2 e_3$		.2179	.0454
		$k_{3(x)}$			.4394
		Σe_3^2			.1931

It is possible to utilize a scheme of successive division if all these correlations are desired when there are more than two predicted variables. By divisions we compute in turn $\rho_{i(x)}$, $\rho_{y_i y_j}$, $\rho_{y_i y_j}$ and $\rho_{y_i y_j}$ from the multiple correlation matrix and $\kappa_{i(x)}$ $\rho_{e_i y_j}$, $\rho_{y_i e_j}$, $\rho_{e_i e_j}$ from the multiple alienation matrix for each i, j. The computational scheme is illustrated in Table IV where the correlations used are the sample correlations of Table III. The calculations from the multiple correlation matrix are presented in Table IVa and those from the multiple alienation matrix in Table IVb.

In Table IVa the multiple correlation matrix is first entered on the third of each three lines. The square root of each diagonal term is then extracted to give the multiple correlation coefficients. The value of $r_{i(x)}$ is then locked in the machine as a divisor and it is divided, in turn, into $\Sigma y_1 y_2$, $\Sigma y_1 y_3$ to get $r_{y_1 y_2}$ and $r_{y_1 y_3}$. Then $r_{2(x)}$ is used as a divisor by division into $r_{y_1 y_2}$ to get $r_{y_1 y_2}$, into $\Sigma y_1 y_2$ to get $r_{y_1 y_2}$, and into $\Sigma y_2 y_3$ to get $r_{y_2 y_3}$. Finally $r_{3(x)}$ is divided into $r_{y_1 y_3}$ to get $r_{y_1 y_3}$, into $\Sigma y_1 y_3$ to get $r_{y_1 y_3}$, into $r_{y_2 y_3}$ to get $r_{y_2 y_3}$ and into $\Sigma y_2 y_3$ to get $r_{y_2 y_3}$. A check on these divisions can be made, if desired, by dividing $r_{y_1 y_2}$ by $r_{1(x)}$ to get $r_{y_1 y_2}$, $r_{y_1 y_3}$ by $r_{1(x)}$ to get $r_{y_1 y_3}$, and $r_{y_2 y_3}$ by $r_{2(x)}$ to get $r_{y_2 y_3}$.

Table IVb is treated in a similar manner.

This technique is immediately applicable to the case of many predicted variables.

REFERENCES

[1] M. Ezekiel, *Methods of Correlation Analysis*, Second Edition, Wiley, 1942.

[2] A. C. Aitken, "On the evaluation of determinants, the formation of their adjugates, etc.," *Edinb. Math. Soc. Proc.*, Series 2, Vol. 3(1933), pp. 207–219.

[3] P. S. Dwyer, "The evaluation of determinants," *Psychometrika*, Vol. 6(1941), pp. 191–204.

[4] A. C. Aitken, *Determinants and Matrices*, Second Edition, Oliver and Boyd, Edinburgh, 1942.

[5] F. V. Waugh and P. S. Dwyer, "Compact computation of the inverse of a matrix," *Annals of Math. Stat.*, Vol. 16(1945), pp. 359–371.

[6] P. S. Dwyer, "A matrix presentation of least squares and correlation theory, etc.," *Annals of Math. Stat.*, Vol. 15(1944), pp. 82–89.

[7] P. S. Dwyer, "The square root method and its use in correlation and regression," *Am. Stat. Assn. Jour.*, Vol. 40(1945), pp. 493–503.

[8] D. B. Duncan and J. F. Kenney, *On the Solution of Normal Equations and Related Topics*, Edwards Brothers, Ann Arbor, 1946.

[9] A. C. Aitken, "The evaluation of a certain triple product matrix," *Roy. Soc. of Edinb. Proc.*, Vol. 57(1937), pp. 172–181.

[10] H. C. Carver, *Anthropometric Data*, Edwards Brothers, Ann Arbor, 1941.

[11] Irving Lorge, "The computation of Hotelling canonical correlations," *Proceedings of Educational Research Forum*, Endicott, N. Y., Aug. 26–31, 1940, pp. 68–74.

INVERSION FORMULAS IN NORMAL VARIABLE MAPPING

By John Riordan

Bell Telephone Laboratories, New York

1. Summary. The two inversion formulas considered here arise from study of G. A. Campbell's work on the Poisson summation, which is described more fully in the introduction and in the main consists of finding a function or mapping of a variable connected with the summation in terms of a normal (Gaussian) variable g. More generally, this last is a process often called "normalization of the variable" and associated with the names of E. A. Cornish and R. A. Fisher. The mapping is two-way and the main inversion formula determines co-efficients for one way from those for the other, both sets of coefficients being descriptive of their mappings. More precisely if x is a given variable, g a Gaussian variable, y a parameter of the mapping, and the two mappings are

$$x = g + \sum_1^\infty G_n(g)\, y^n/n!,$$

$$g = x + \sum_1^\infty X_n(x) y^n/n!,$$

the formula expresses $G_n(x)$ in terms of $X_i(x)$, $i \leq n$, and vice versa.

The second formula is more particularly related to the Poisson summation and relates coefficients $p_n \equiv p_n(g)$ and $q_n \equiv q_n(g)$ in the pair of equations

$$a = c \sum_0^\infty q_n c^{-\frac{1}{2}n}/n!$$

$$c = a \sum_0^\infty p_n a^{-\frac{1}{2}n}/n!$$

Both formulas, which are necessarily elaborate, are given concise expression by the use of the multi-variable polynomials of E. T. Bell.

2. Introduction. In 1923, in a paper little known in statistical circles, G. A. Campbell [2] gave as the basis for his extensive tabulation of the Poisson summation an asymptotic series expressing the average a in terms of a normal variable g, corresponding to the probability of at least c occurrences, and c itself. That is to say, he associated with the Poisson summation

$$P(a, c) = \sum_c^\infty e^{-a} a^x/x!$$

a normal variable g, defined by

$$P(a, c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{g} e^{-x^2/2}\, dx$$

and inverted the summation (which, as is well known, is equivalent to the incomplete Gamma function ratio) to give a series for a in terms of g and c. The series, which is carried to 11 terms, starts as follows:

$$a \sim c\left[1 + gc^{-1/2} + \frac{g^2 - 1}{3}\, c^{-1} + \frac{g^3 - 7g}{36}\, c^{-3/2} + \cdots\right]$$

If $x = (a - c)\, c^{-1/2}$ is introduced, this becomes

$$x \sim g + \frac{g^2 - 1}{3}\, c^{-1/2} + \frac{g^3 - 7g}{36}\, c^{-1} + \cdots$$

and x is seen to be, like g, a standardized variable of mean 0, variance 1.

It seems to have gone unnoticed that this result includes the χ^2 distribution through the transformation: $2a = X^2$, $2c = n$ and it has been rediscovered by A. M. Peiser [7] (4 terms) and by Goldberg and Levine [4] (6 terms).

It is possible also to express c in terms of a and g, and a formula of this kind with fewer terms which appears in a footnote in Campbell's paper is as follows:

$$c \sim a\left[1 - ga^{-1/2} + \frac{g^2 + 2}{6}\, a^{-1} + \frac{g^3 + 2g}{72}\, a^{-3/2} + \cdots\right]$$

Finally there is a third possibility of expressing g in terms of the remaining variables, preferably x and c; though unnoticed by Campbell this has since been brought to prominence by Cornish and Fisher [3], Hotelling and Frankel [5] and Kendall [6].

The idea behind the first expansion appears most clearly in the second form and is that for c large the variable x behaves nearly like g. The third possibility reverses this expansion and gives a function of x and c which behaves like g; hence if this function is first evaluated, reference to the normal integral table gives an immediate evaluation of the probabilities in question. Put in another way, the expansion widens the scope of the normal integral table and for this reason has been called "normalization" of the variable (but this term seems preempted by its use in another sense for orthogonal functions, and has been replaced in the title by normal variable mapping).

From the point of view of statistical theory, the three expressions are different versions of one relationship, which suggests that there should be general rules for transforming a series of one type into that of another. The two inversion formulas given below supply these rules in what appears to be as compact a form as the problem allows. It will be noted that the proofs given suppose convergent series, a case which leads to clarity and brevity and is interesting in itself. Applied to Campbell's series, they give the known results so far as the latter go, but of course for other asymptotic series they need independent verifications.

3. First Inversion Formula. This relates coefficients in series like Campbell's first and its reverse as in Cornish and Fisher. More precisely

If $G_1(g)$, $G_2(g)$ $\cdots$ are assigned polynomials and if

$$x = g + \sum_{n=1}^{\infty} G_n(g)y^n/n!, \tag{1}$$

defines x in terms of g and a parameter y, then

$$g = x + \sum_{1}^{\infty} X_n(x)y^n/n!, \tag{2}$$

where

$$-X_n(x) = Y_n(aG_1(x), aG_2(x), \cdots, aG_n(x)), \tag{3}$$

TABLE 1

Bell Polynomials Y_n (fg_1, fg_2 $\cdots$ fg_n)

$$\begin{aligned}
Y_1 &= f_1g_1\\
Y_2 &= f_1g_2 + f_2g_1^2\\
Y_3 &= f_1g_3 + f_2(3g_2g_1) + f_3g_1^3\\
Y_4 &= f_1g_4 + f_2(4g_3g_1 + 3g_2^2) + f_3(6g_2g_1^2) + f_4g_1^4\\
Y_5 &= f_1g_5 + f_2(5g_4g_1 + 10g_3g_2) + f_3\,(10g_3g_1^2 + 15g_2^2g_1)\\
&\quad + f_4\,(10g_2g_1^3) + f_5g_1^5\\
Y_6 &= f_1g_6 + f_2(6g_5g_1 + 15g_4g_2 + 10g_3^2)\\
&\quad + f_3(15g_4g_1^2 + 60g_3g_2g_1 + 15g_2^3)\\
&\quad + f_4(20g_3g_1^3 + 45g_2^2g_1^2) + f_5(15g_2g_1^4) + f_6g_1^6\\
Y_7 &= f_1g_7 + f_2(7g_6g_1 + 21g_5g_2 + 35g_4g_3)\\
&\quad + f_3(21g_5g_1^2 + 105g_4g_2g_1 + 70g_3^2g_1 + 105g_3g_2^2)\\
&\quad + f_4(35g_4g_1^3 + 210g_3g_2g_1^2 + 105g_2^3g_1)\\
&\quad + f_5(35g_3g_1^4 + 105g_2^2g_1^3) + f_6(21g_2g_1^5) + f_7g_1^7\\
Y_8 &= f_1g_8 + f_2(8g_7g_1 + 28g_6g_2 + 56g_5g_3 + 35g_4^2)\\
&\quad + f_3(28g_6g_1^2 + 168g_5g_2g_1 + 280g_4g_3g_1 + 210g_4g_2^2 + 280g_3^2g_2)\\
&\quad + f_4(56g_5g_1^3 + 420g_4g_2g_1^2 + 280g_3^2g_1^2 + 840g_3g_2^2g_1 + 105g_2^4)\\
&\quad + f_5(70g_4g_1^4 + 560g_3g_2g_1^3 + 420g_2^3g_1^2)\\
&\quad + f_6(56g_3g_1^5 + 210g_2^2g_1^4) + f_7(28g_2g_1^6) + f_8g_1^8
\end{aligned}$$

Y_n being the multivariable polynomial of E. T. Bell [1], *in the variables $G_1(x)$ to $G_n(x)$ and the symbolic variable a which is such that*

$$a^i \equiv a_i = (-D)^{i-1}, \qquad D = d/dx,$$

with differentiations on all products of $G_1(x)$ to $G_n(x)$ associated with it in the polynomial.

Note the symmetry of x and g, which allows the transformation to go either way, the inverse of (3) being

$$-G_n(g) = Y_n(aX_1(g), aX_2(g) \cdots, aX_n(g)) \tag{4}$$

Table I gives explicit expressions for polynomials Y_1 to Y_8. It will be noted that the number of terms in Y_n is the number of partitions of n and that f_i, the

variable replacing a_i in the table, is associated with terms corresponding to partitions with i parts; that is to say, if $Y_{n,i}$ designates such terms

$$Y_n = \sum_1^n f_i Y_{n,i}$$

The verification or extension of the table may be accomplished by the formulas and relations given by Bell (l.c.) or more directly by those modifications of Bell given by myself in [8].

The first few instances of (3), dropping the common variable x for brevity, may be read off from Table I (with appropriate changes of notation and interpretation of a_i) as follows:

$$-X_1 = G_1$$
$$-X_2 = G_2 - D(G_1^2)$$
$$-X_3 = G_3 - 3D(G_2G_1) + D^2(G_1^3)$$
$$-X_4 = G_4 - 4D(G_3G_1) - 3D(G_2^2) + 6D^2(G_2G_1^2) - D^3(G_1^4)$$

Applied to Campbell's first formula in its second form with $y = c^{-1/2}$ and

$$G_1(x) = (x^2 - 1)/3, \qquad G_3(x) = (-6x^4 - 14x^2 + 32)/270,$$
$$G_2(x) = (x^3 - 7x)/18, \qquad G_4(x) = (9x^5 + 256x^3 - 433x)/1680,$$

these show e.g.

$$-X_2 = \frac{x^3 - 7x}{18} - \frac{2(x^2 - 1)}{3} \cdot \frac{2x}{3} = \frac{-7x^3 + x}{18},$$

and similarly for the others, resulting in

$$X_1 = -(x^2 - 1)/3$$
$$X_2 = (7x^3 - x)/18$$
$$X_3 = -(219x^4 - 14x^2 - 13)/270$$
$$X_4 = (3993x^5 - 152x^3 + 119x)/1680$$

These determine a calculation formula for the Poisson summation, which is a refinement of the normal approximation. That is to say

$$P(a, c) = \Phi(g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{g} e^{-t^2/2}\, dt$$

with

$$g = x - \frac{x^2 - 1}{3\sqrt{c}} + \frac{7x^3 - x}{36c} - \frac{219x^4 - 14x^2 - 13}{1620c\sqrt{c}} + \frac{3993x^5 - 152x^3 + 119x}{40320c^2} - \cdots$$

and $x = (a - c)/\sqrt{c}$.

For the t-variate, the formula is applied in the reverse direction since Hotelling and Frankel supply the first four values of X_n, that is, in present notation, the series

$$g \sim x - \frac{x^3 + x}{4}\, y + \frac{13x^5 + 8x^3 + 3x}{48}\,\frac{y^2}{2} - \frac{35x^7 + 19x^5 + x^3 - 15x}{64}\,\frac{y^3}{6} + \frac{6271x^9 + 3224x^7 - 102x^5 - 1680x^3 - 945x}{3840}\,\frac{y^4}{24} + \cdots$$

The reversed series (obtained by (4)) is

$$x \sim g + \frac{g^3 + g}{4}\, y + \frac{5g^5 + 16g^3 + 3g}{48}\,\frac{y^2}{2} + \frac{3g^7 + 19g^5 + 17g^3 - 15g}{64}\,\frac{y^3}{6} + \frac{79g^9 + 776g^7 + 1482g^5 - 1920g^3 - 945g}{3840}\,\frac{y^4}{24} + \cdots$$

The first three terms are checked by Goldberg and Levine (l.c.).

Another application worth noting is to the formulas of Cornish and Fisher which give $G_i(g)$ and $X_i(x)$ in terms of the relative cumulants of the distribution; to save space these are omitted.

The derivation of the formula may be indicated most easily by Lagrange's formula for the expansion of one function in powers of another in the following form[1]:

Let C be a contour in the complex z plane enclosing the point $z = x$, and let $f(z)$ and $\phi(z)$ be analytic on and inside C. Let y be such that $|\, y\phi(z)\,| < |\, z - x\,|$ when z is on C, and g be that root of the equation:

$$g = x + y\phi(g) \tag{5}$$

which lies inside C. Then

$$f(g) = \frac{1}{2\pi i}\int_C f(z)\,\frac{d}{dz}\{\log[z - x - y\phi(z)]\}\,dz = f(x) + \sum_1^\infty X_n^*(x)y^n/n! \tag{6}$$

where

$$X_n^*(x) = \frac{d^{n-1}}{dx^{n-1}}[f'(x)(\phi(x))^n] \tag{7}$$

The contour integral in (6) appears, slightly disguised, as a problem in Whittaker and Watson [*Modern Analysis*, Cambridge, 1920, p. 149]. The evaluation (7) is given for completeness, though no use is made of it in this section, the derivation proceeding directly from (6).

First notice that by (1) and (5)

$$-y\phi(g) = \sum_1^\infty G_n(g)y^n/n!,$$

[1] The author owes the suggestion for this to S. O. Rice, who also simplified the derivation of the second inversion formula given later.

so that the logarithm in (6) may be written

$$\log (z - x + \sum_1^\infty G_n(z)y^n/n!),$$

or

$$\log (z - x) + \log \left[1 + \sum_1^\infty G_n(z)(z - x)^{-1}y^n/n!\right],$$

or

$$\log (z - x) + \log \exp by, \tag{8}$$

with b a symbolic variable such that

$$b^0 \equiv b_0 = 1$$
$$b^n \equiv b_n = G_n(z)(z - x)^{-1}.$$

Now if

$$\log (\exp by) = B_1y + B_2y^2/2! + \cdots, \tag{9}$$
$$= \exp By,$$

B being another symbolic variable, $B_0 = 0$, $B^n \equiv B_n$, it follows from equation (5) of [8] that

$$B_n = [D_y^n \log (\exp by)]_{y=0}, \qquad D_y = d/dy,$$
$$= Y_n(\beta b_1, \beta b_2, \cdots \beta b_n)$$
$$= \sum_1^n \beta_i Y_{n,i}(b_1, b_2, \cdots b_n),$$

with $\beta_i = (-)^{i-1}(i - 1)!$ and $Y_{n,i}$ the part of polynomial Y_n having i parts, as defined above. Moreover, each factor b_k of terms in $Y_{n,i}$ contributes $G_k(z)(z - x)^{-1}$ so that

$$B_n = \sum_1^n \beta_i(z - x)^{-i} Y_{n,i}(G_1(z), G_2(z) \cdots G_n(z)) \tag{11}$$

Then, by (5)

$$f(g) = \frac{1}{2\pi i}\int_C f(z)\left(\frac{1}{z - x} + \frac{d}{dz}\exp By\right)dz$$
$$= f(x) - \frac{1}{2\pi i}\int_C f'(z)\exp By\, dz$$
$$= f(x) - \frac{1}{2\pi i}\int_C f'(z)\sum_1^\infty \frac{y^n}{n!}\sum_1^n \frac{\beta_i Y_{n,i}(G_1 \cdots G_n)}{(z - x)^i}\, dz$$

$$= f(x) - \sum_1^\infty \frac{y^n}{n!} \sum_{i=1}^n \int_C \frac{(-)^{i-1}(i-1)!}{2\pi i(z-x)^i} Y_{n,i}(G_1(z) \cdots G_n(z)) f'(z)\, dz$$

$$= f(x) - \sum_1^\infty \frac{y^n}{n!} \sum_1^n (-D)^{i-1}[f'(x) Y_{n,i}(G_1(x) \cdots G_n(x))]$$

with $D = d/dx$. The evaluation in the last line is by the Cauchy formula for derivates; the second line is derived by an integration by parts.

Equation (4) follows from this and the substitution $f(g) = g$.

4. Second Inversion Formula. This gives the interrelations of coefficients of series like the two Campbell series mentioned in the introduction. It runs as follows:

If $q_1(g)$, $q_2(g)$ $\cdots$ are given polynomials and if

$$a = c \sum_0^\infty \frac{q_n(g) c^{-\frac{1}{2}n}}{n!} \tag{12}$$

defines a in terms of g and a parameter c; then

$$c = a \sum_0^\infty \frac{p_n(g) a^{-\frac{1}{2}n}}{n!} \tag{13}$$

where

$$-p_n(g) = Y_n(\alpha q_1(g),\ \alpha q_2(g),\ \cdots,\ \alpha q_n(g)) \tag{14}$$

with $\alpha^1 \equiv \alpha_1 = 1$; $\alpha^i \equiv \alpha_i = (n-4)(n-6) \cdots (n-2i)2^{-i+1}$

Equation (14) is formally similar to (3) and by symmetry as before, $q_n(g)$ is readily expressible as a Y_n polynomial in $p_1(g)$ to $p_n(g)$.

The first five instances of (14), dropping the argument for brevity, are

$$-p_1 = q_1$$

$$-p_2 = q_2 - q_1^2$$

$$-p_3 = q_3 - \tfrac{3}{2} q_2 q_1 + \tfrac{3}{4} q_1^3$$

$$-p_4 = q_4$$

$$-p_5 = q_5 + \tfrac{5}{2}(q_4 q_1 + 2q_3 q_2) - \tfrac{5}{4}(2q_3 q_1^2 + 3q_2^2 q_1) + \tfrac{15}{4} q_2 q_1^3 - \tfrac{15}{16} q_1^5$$

Applied to Campbell's first series where

$$q_1(g) = g \qquad q_3(g) = (g^3 - 7g)/6$$

$$q_2(g) = \tfrac{2}{3}(g^2 - 1) \qquad q_4(g) = (-12g^4 - 28g^2 + 64)/135$$

$$q_5(g) = (36g^5 + 1024g^3 - 1732g)/1296$$

these show that

$$p_1(g) = -g \qquad p_3(g) = (g^3 + 2g)/12$$
$$p_2(g) = (g^2 + 2)/3 \qquad p_4(g) = (12g^4 + 28g^2 - 64)/135$$
$$p_5(g) = (207g^5 + 2596g^3 - 6148g)/1296$$

The proof of (14) is as follows. First, for brevity introduce symbolic variables p and q with the usual interpretation $p^n \equiv p_n(g)$, $q^n \equiv q_n(g)$ so that (12) and (13) read

$$a = c \exp q\, c^{-\frac{1}{2}}$$
$$c = a \exp p\, a^{-\frac{1}{2}}$$

Now write $a = 1/x^2$, $c = 1/y^2$ changing these to

$$x = y\,(\exp qy)^{-\frac{1}{2}}$$
$$y = x\,(\exp px)^{-\frac{1}{2}}$$

and note that

$$x^2y^{-2} = (\exp qy)^{-1} = \exp px \tag{15}$$

which shows that p_n is the coefficient of $x^n/n!$ in the expansion in powers of x of $(\exp qy)^{-1}$. Lagrange's formula gives at once ($D = d/dy$):

$$f(y) = \sum_1^\infty \frac{x^n}{n!} D^{n-1}[f'(g)(\exp qy)^{\frac{1}{2}n}]_{y=0} \tag{16}$$

so that

$$\begin{aligned}(\exp qy)^{-1} &= \sum_1^\infty \frac{x^n}{n!} D^{n-1}[-(\exp qy)^{\frac{1}{2}(n-4)} D(\exp qy)]_{y=0} \\ &= \sum_1^\infty \frac{x^n}{n!} D^{n-1}\left[-\frac{2}{n-2} D(\exp qy)^{\frac{1}{2}(n-2)}\right]_{y=0} \\ &= \sum_1^\infty \frac{x^n}{n!}\left(\frac{-2}{n-2}\right)[D^n(\exp qy)^{\frac{1}{2}(n-2)}]_{y=0}\end{aligned}$$

or

$$\begin{aligned}-p_n &= \frac{2}{n-2}[D^n(\exp qy)^{\frac{1}{2}(n-2)}]_{y=0} \\ &= Y_n(\alpha q_1, \alpha q_2, \cdots, \alpha q_n)\end{aligned} \tag{17}$$

with α_i as in (14), by equation (5) of [8].

BIBLIOGRAPHY

[1] Bell, E. T., "Exponential polynomials," *Annals of Math.* Vol. *35*, (1934), pp. 258–277.
[2] Campbell, G. A., "Probability curves showing Poisson's exponential summation,"

Bell System Tech Jl. Vol. *2*, (1923), pp. 95–113; Collected Papers, New York, 1937, pp. 224–242.

[3] Cornish, E. A. and Fisher, R. A., "Moments and cumulants in the specification of distributions," *Revue de l'Institut Intern. de Stat.* Vol. *4*, (1937), pp. 1–14.

[4] Goldberg, H. and Levine, H., "Approximate formulas for the percentage points and normalization of t and χ^2," *Annals of Math. Stat.* Vol. *17*, (1946), pp. 216–225.

[5] Hotelling, H. and Frankel, L. R., "The transformation of statistics to simplify their distribution," *Annals of Math. Stat.* Vol. *9*, (1938), pp. 87–96.

[6] Kendall, M. G., *The Advanced Theory of Statistics* I, London. 1943.

[7] Peiser, A. M., "Asymptotic formulas for significance levels of certain distributions," *Annals of Math. Stat.* Vol. *14*, (1943), pp. 56–62.

[8] Riordan, John, "Derivatives of Composite Functions," *Bull. Am. Math. Soc.* Vol. *52*, (1946), pp. 664–667.

ON THE DETERMINATION OF OPTIMUM PROBABILITIES IN SAMPLING

By Morris H. Hansen and William N. Hurwitz

Bureau of the Census

1. Summary. In a previous paper [2] it was shown that it is sometimes profitable to select sampling units with probability proportionate to size of the unit. This note indicates a method of determining the probabilities of selection which minimize the variance of the sample estimate at a fixed cost. Some approximations that have practical applications are given.

2. Introduction. Neyman has shown that it is possible to reduce the sampling variance of an estimate by dividing a population into sub-populations (called strata) and varying the proportions of units included in the sample from stratum to stratum [1]. His treatment presumed that the units within any stratum would be drawn with equal probability. In many practical sampling problems, the use of constant probabilities is neither necessary nor desirable. Not only is it possible to obtain unbiased or consistent estimates with varying probabilities of selection of the sampling units, but also it is possible to reduce the variance of sample estimates by appropriate use of this device.

It has been shown [2] that in a subsampling system, the selection of primary units with probabilities proportionate to the number of elements included in the primary unit may bring about marked reductions in sampling variances over sampling with equal probabilities. In this note, we shall indicate a method of determining the optimum probabilities under certain conditions, and also some approximations to the optima that have practical applications.

By optimum probabilities, we mean the set of probabilities of selection that will minimize the variance for a fixed cost of obtaining sample results, or alternatively that will minimize the cost for a fixed sampling error.

3. Optimum probability with a subsampling system. Consider, for example, the simple subsampling system where primary units are first drawn for inclusion in the sample and then a sample of elements is drawn from the selected primary units. We shall suppose, for simplicity of notation, that the sampling is done without stratification. The conclusions indicated below will be similar if stratified sampling is used, and they will hold even if only one unit is drawn from each stratum. Suppose that a population contains M primary units, and that the sampling of primary units is to be done with replacement. Sampling with replacement is assumed in order to simplify the mathematics. We wish to estimate the ratio

$$\frac{X}{Y} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} X_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N_i} Y_{ij}}$$

where X_{ij} and Y_{ij} are the values of two characteristics of the jth element within the ith primary unit, and N_i is the number of elements in the ith primary unit. A consistent estimate of X/Y is given by

$$r = \frac{\sum_{i=1}^{m} \frac{N_i}{P_i} \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}}{\sum_{i=1}^{m} \frac{N_i}{P_i} \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}} \tag{1}$$

where

P_i = The probability of selecting the ith primary unit on a single draw.

n_i = The total number of elements included in the sample from the ith unit if it is drawn. If a particular unit happens to be included in the sample more than once the subsampling will be independently carried through each time it is drawn.

m = The total number of primary units included in the sample.

It will be assumed that a self-weighting sample is to be used, i.e., that although the probabilities of selecting primary units will vary, the subsampling rate within the ith selected primary unit, $\frac{n_i}{N_i}$, will be such that $P_i \frac{n_i}{N_i} = k$. Note that, with this condition, k is the probability that an element will be included in the sample by making a single draw of a primary unit, and by carrying out the specified subsampling within the selected primary unit. It follows that mkN is the expected total number of elements included in a sample of primary units, where

$$N = \sum_{i=1}^{M} N_i.$$

The method can be extended to cover situations where other conditions are imposed.

We shall express the variance of r in terms of P_i, m, and k, and also express the cost in terms of these same quantities. The optimum values of P_i, m, and k will then be determined.

The variance of the sample estimate. To terms of order $1/m$ of the Taylor expansion of a ratio, the sampling variance of the estimate (1) is approximately

$$\sigma_r^2 \doteq \frac{\sum_{i=1}^{M} \frac{N_i^2}{P_i} \Delta_i^2 + \sum_{i=1}^{M} \frac{N_i^2}{P_i} \frac{N_i - n_i}{N_i n_i} \sigma_i^2}{mY^2} \tag{2}$$

where

$$\Delta_i^2 = \bar{Y}_i^2 \left(\frac{\bar{X}_i}{\bar{Y}_i} - \frac{X}{Y} \right)^2, \qquad \bar{Y}_i = \frac{\sum_{j=1}^{N_i} Y_{ij}}{N_i}, \qquad \bar{X}_i = \frac{\sum_{j=1}^{N_i} X_{ij}}{N_i},$$

$$\sigma_i^2 = \sigma_{ix}^2 + \frac{\bar{X}^2}{\bar{Y}^2}\sigma_{iy}^2 - 2\frac{\bar{X}}{\bar{Y}}\sigma_{ixy},$$

$$\sigma_{ix}^2 = \frac{\sum_{j=1}^{N_i}(X_{ij} - \bar{X}_i)^2}{N_i - 1},$$

$$\sigma_{iy}^2 = \frac{\sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i)^2}{N_i - 1},$$

$$\sigma_{ixy} = \frac{\sum_{j=1}^{N_i} X_{ij} - \bar{X}_i)(Y_{ij} - \bar{Y}_i)}{N_i - 1}.$$

The cost function. Now suppose that the total cost of the sampling procedure involves a fixed cost attached to each primary unit included in the sample, a cost of listing the elements within each selected primary unit (this listing may be necessary in order to draw a subsample), and a cost of obtaining information from each of the elements selected for inclusion in the sample. Under these circumstances the total expected cost of the survey will be:

$$C = C_1 m + C_2 m \sum_{i=1}^{M} P_i N_i + C_3 mkN \tag{3}$$

where

C_1 = The fixed cost per primary unit,
C_2 = The cost of listing one element in a selected primary unit and other costs that vary with the number of elements to be listed,
C_3 = The cost of obtaining the required information from one element in the sample,
$\sum_{i=1}^{M} P_i N_i$ = Expected number of elements in the sample per primary unit in the sample,
mk = The over-all sampling ratio, and
$N = \sum_{i=1}^{M} N_i$ = The total number of elements in the population.

It will be noted that although the values of P_i and m may be fixed in advance, the number of elements to be listed, $\sum_{i=1}^{m} N_i$, remains a chance variable. It is for this reason that we consider the expected cost rather than the actual cost.

The optimum values of P_i, m, and k. The values of P_i, m, and k which minimize the variance (2) subject to the conditions that:

$$C \text{ is fixed}, \qquad \frac{n_i}{N_i} P_i = k, \qquad \sum_{i=1}^{M} P_i = 1,$$

are given by

$$P_i = \frac{\sqrt{\frac{N_i^2 \delta_i}{C_1 + C_2 N_i}}}{\sum_{i=1}^{M} \sqrt{\frac{N_i^2 \delta_i}{C_1 + C_2 N_i}}}, \tag{4}$$

$$k = \frac{\sqrt{\frac{\sum_{i=1}^{M} N_i \sigma_i^2}{N}}}{\sum_{i=1}^{M} \sqrt{\frac{N_i^2 \delta_i}{C_1 + C_2 N_i}} C_3}, \tag{5}$$

$$m = \frac{C}{C_1 + C_2 \sum_{i=1}^{M} P_i N_i + C_3 k N}, \tag{6}$$

where

$$\delta_i = \Delta_i^2 - \frac{\sigma_i^2}{N_i}.$$

Ordinarily δ_i will be positive although it will often be found to be negative for some i. For a great many populations, such negative values can be avoided by classifying the primary units into size groups or other significant groups and then requiring that the probability of selection be P_α for every primary unit in the α-th group.

In actual practice, however, in advance of designing a sample one does not have the data to compute the optima and uses methods of approximating the optimum probabilities. Methods of approximating the optimum probabilities are given below.

4. Some rules for approximating the optimum probabilities. In another paper [2] considerations were presented from which it follows that δ_i tends to decrease with increasing size of unit, but seldom as fast as the size of unit increases. The rate of decrease is often small relative to the increase in N_i, and empirical data for a number of problems indicate that even the assumption of δ_i being fairly constant with increasing size of unit may not lead one far astray from the optimum probabilities. Under this assumption ($\delta_i = \delta$ for all i) the probabilities depend only on N_i, C_1, and C_2, and lead to the following results:

(a) When $C_1 > 0$ and $C_2 = 0$, probability proportionate to size will be the optimum.

(b) When $C_1 = 0$ and $C_2 > 0$, probability proportionate to the square root of the size will be the optimum.

If we go to the other extreme (extreme not in terms of mathematically possible values but in terms of most practical populations), and assume that δ_i decreases at the same rate that N_i increases, the results would be:

(a) When $C_1 > 0$ and $C_2 = 0$, probability proportionate to the square root of the size will be the optimum.

(b) When $C_1 = 0$ and $C_2 > 0$, equal probability will be the optimum.

The minimum is broad in the neighborhood of the optimum and the results for either of these extremes and the values in between often will give results reasonably close to the minimum. This leads to the following useful approximations:

(a) When $C_2 \Sigma P_i N_i$, the expected cost per primary unit of listing and related operations, is small in relation to C_1, the fixed cost per primary unit, the optimum probabilities will be between probability proportionate to size and probability proportionate to the square root of size, and either of these will be reasonably close to the optimum.

(b) When C_1 is small compared to $C_2 \Sigma P_i N_i$, the optimum probability will be between equal probability and probability proportionate to the square root of size, and either of these will be reasonably close to the optimum.

(c) When both C_1 and $C_2 \Sigma P_i N_i$ are of significant size, i.e., when the costs vary substantially both with the number of primary units in the sample and the size of the units, then probability proportionate to the square root of the size will be a reasonably good approximation to the optimum.

(d) When units of small size are used and all of the subunits in the selected primary units are included in the sample (that is, there is no subsampling) equal probability is close to the optimum. It should be noted that this rule does not follow directly from the above analysis based on subsampling, but from a separate analysis in which no subsampling is involved.

For whatever system of probabilities is used, and with the cost function given by (3), the optimum value of k is given by:

$$k = \sqrt{\frac{\sum_{i=1}^{M} N_i \sigma_i^2 \left(C_1 + C_2 \sum_{j=1}^{M} P_j N_j \right)}{C_3 N \left(\sum_{i=1}^{M} \frac{N_i^2 \Delta_i^2}{P_i} - \sum_{i=1}^{M} \frac{N_i \sigma_i^2}{P_i} \right)}},$$

which can be approximated, in application, from prior experience or preliminary studies. The corresponding optimum value for m is obtained by substitution in the cost function.

The above results should not be accepted, of course, as the optima for every cost function or every sampling system. Either past experimental data may be available or pilot tests made to determine the cost function and the appropriate approximations that should be used in various practical situations.

An illustration. An illustration may be of interest. A characteristic published for city blocks in the 1940 Census of Housing is the number of dwelling units that are in need of major repairs or that lack a private bath. Suppose we

were sampling to estimate the proportion of the dwelling units having this characteristic for the Bronx in New York City, at the time of the 1940 Census. Let us assume that once we selected a system of probabilities we used the optimum numbers of blocks and the optimum sampling ratios appropriate to these probabilities, that is, the optimum values of k and m. For each of several cost functions the following Table 1 shows the sampling variances of each system, rela-

TABLE 1

Unit costs			Average cost per primary unit of listing and related operations ($C_2\Sigma P_i N_i$)			Variances relative to equal probability		
C_1	C_2	C_3	Equal probability	Probability proportionate to square root of size	Probability proportionate to size	Equal probability	Probability proportionate to square root of size	Probability proportionate to size
5	.10	1	13.49	21.15	27.63	100	92	104
5	.05	1	6.75	10.58	13.82	100	88	97
5	.02	1	2.70	4.23	5.53	100	83	87
5	0	1	0	0	0	100	75	73
2	.10	1	13.49	21.15	27.63	100	96	111
2	.05	1	6.75	10.58	13.82	100	93	106
2	.02	1	2.70	4.23	5.53	100	90	97
2	0	1	0	0	0	100	79	77
1	.10	1	13.49	21.15	27.63	100	97	114
1	.05	1	6.75	10.58	13.82	100	96	110
1	.02	1	2.70	4.23	5.53	100	93	103
1	0	1	0	0	0	100	82	81
0	.10	1	13.49	21.15	27.63	100	99	117
0	.05	1	6.75	10.58	13.82	100	99	115
0	.02	1	2.70	4.23	5.53	100	99	113

tive to the variance of sampling with equal probability. It also shows values of $C_2\Sigma P_i N_i$ for comparison with C_1.

Some of the costs given in the table do not have unreasonable relationships in terms of the situations encountered in practice in various types of jobs. The comparisons are not affected by the absolute magnitudes of the costs but only by their relative magnitudes. The results are consistent with the rough rules of thumb given above. It is worth noting that in each of the above instances probability proportionate to the square root of the size yields a comparatively low variance.

5. Sampling with or without replacement. In this paper the sampling with varying probabilities was assumed to be carried out with replacement which ordinarily would not be advisable in practice. When sampling is done without replacement the optimum probabilities and their approximations will be about the same as for sampling with replacement in at least those instances where the proportion of the population in the sample is small. Further investigation is needed for large sampling rates.

6. Conclusion. In summary, it is not essential and may not be desirable to give each element in the population (or stratum) the same chance of being drawn in order to avoid bias or to have a consistent estimate. Estimate (1) is a consistent estimate no matter what probabilities of selection are assigned to these units. The use of variable probabilities of selection is another device to be added to those already in the literature, such as stratification and efficient methods of estimation, which make it possible to achieve the objectives of a sample survey at reduced costs. Reference [2] gives another illustration of reductions in sampling variance achieved through the use of varying probabilities in accordance with the rules suggested above for approximating the optimum probabilities.

REFERENCES

[1] JERZY NEYMAN, "On the two different aspects of the representative method of purposive selection," *Roy. Stat. Soc. Jour.*, New Series, Vol. 97 (1934), pp. 558–606.

[2] MORRIS H. HANSEN AND WILLIAM N. HURWITZ, "On the theory of sampling from finite populations," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 333–362

A SOLUTION TO THE PROBLEM OF OPTIMUM CLASSIFICATION

By P. G. Hoel and R. P. Peterson

University of California, Los Angeles

1. Summary. By means of a general theorem, the space of the variables of classification is separated into population regions such that the probability of a correct classification is maximized. The theorem holds for any number of populations and variables but requires a knowledge of population parameters and probabilities. A second theorem yields a large sample criterion for determining an optimum set of estimates for the unknown parameters. The two theorems combine to yield a large sample solution to the problem of how best to discriminate between two or more populations.

2. Introduction. There are essentially two basic problems in discriminant analysis. The first problem is to test whether the populations differ, since it would be futile to attempt a classification if the populations did not differ. The second problem is to find an efficient method for classifying individuals into their proper populations. In this paper, an optimum asymptotic solution of the second problem will be presented.

3. Parameters known. Let $f_i = f_i(x_1, \cdots, x_k)$, $(i = 1, \cdots, r)$ denote the probability density function of population i in the region under consideration. Let $p_i > 0$, $(i = 1, \cdots, r)$, denote the probability that population i will be sampled if a single individual is selected at random from that region, and let R denote the k dimensional Euclidean variable space. Then the desired theorem is the following:

Theorem 1. *If M_i denotes the region in R where $p_i f_i \geq p_j f_j$, $(j = 1, \cdots, r)$, and where $p_i f_i > 0$, then the set of regions M_i, $(i = 1, \cdots, r)$, in which any overlap is assigned to the M_i with the smallest index, will maximize the probability of a correct classification.*

For the purpose of proving this theorem, consider any other set of non-overlapping regions, M_i'. Since the addition to any of the regions M_i of a part of R throughout which all the functions f_i vanish will not affect the probability of a correct classification, there is no loss of generality in assuming that the set of regions M_i' contains the same portion of R as the set of regions M_i does. The relationship between the two sets may be expressed by means of the formulas

$$M_i = \sum_{j=1}^{r} M_{ij} \tag{1}$$

and

$$M_j' = \sum_{i=1}^{r} M_{ij}, \tag{2}$$

where M_{ij} denotes that part of M_i which is contained in M_j'.

Since a sample point that falls in the region M_i will be judged to have come from population i, the probability of the correct classification of a single random sample by means of the set M_i is given by

$$Q = p_1 \int_{M_1} f_1 dE + \cdots + p_r \int_{M_r} f_r dE, \tag{3}$$

where $dE = dx_1 dx_2 \cdots dx_r$. If Q' denotes the probability of the correct classification by means of the set M_i',

$$Q' = p_1 \int_{M_1'} f_1 \, dE + \cdots + p_r \int_{M_r'} f_r \, dE.$$

In the notation of (1) and (2), these probabilities become

$$Q = p_1 \int_{\sum_i M_{1i}} f_1 \, dE + \cdots + p_r \int_{\sum_i M_{ri}} f_r \, dE$$

and

$$Q' = p_1 \int_{\sum_i M_{i1}} f_1 \, dE + \cdots + p_r \int_{\sum_i M_{ir}} f_r \, dE.$$

Now consider the difference $Q - Q'$. It can be expressed in the form

$$\begin{aligned} Q - Q' &= \sum_{i=1}^{r} \sum_{j=1}^{r} \left[p_i \int_{M_{ij}} f_i \, dE - p_j \int_{M_{ij}} f_j \, dE \right] \\ &= \sum_{i=1}^{r} \sum_{j=1}^{r} \int_{M_{ij}} [p_i f_i - p_j f_j] \, dE. \end{aligned}$$

Since M_{ij} is contained in M_i and $p_i f_i \geq p_j f_j$, $(j = 1, \cdots, r)$, holds throughout M_i, it follows that each of these integrals is non-negative; consequently $Q \geq Q'$, which proves the theorem.

This theorem yields a solution to the classification problem only when the f_i are completely specified and the p_i are known.

It will be observed that this theorem is similar to a generalization of a fundamental lemma in the Neyman-Pearson theory of testing hypotheses [1], and to a result by Welch [2].

If the basic weight function in Wald's [3] formulation of the multiple decision problem assumes only the values 0 and 1, corresponding to whether or not a correct classification is made, it will be found that the set of regions M_i will minimize the expected value of the loss in that formulation.

4. Parameters unknown. Since the p_i, as well as the parameters in the f_i, are assumed to be unknown, Q will be a function of such parameters. Let $\theta_1, \cdots, \theta_s$ denote all such parameters, including the p_i. Now let a random sample of size n be taken from the region under consideration and let $\bar{\theta}_1, \cdots, \bar{\theta}_s$ denote a set of

estimates of the parameters based on this sample. Since the total sample will constitute a sample of size n_1 from f_1, n_2 from f_2, etc., where $n = n_1 + \cdots + n_r$, the θ's for f_i will be estimated by means of a sample of size n_i rather than of size n. In the following arguments, it will not be necessary to distinguish between θ's which are estimated by different size samples because the arguments will be based on the order of terms with respect to the size sample and $n_i \sim np_i$ with probability one. Or, more simply, choose all n_i equal.

Let $\bar{M}_i$ correspond to M_i when the parameters are replaced by their sample estimates and let $\bar{Q}$ denote the probability of a correct classification when using the regions $\bar{M}_i$ in place of the regions M_i. Then, from (3),

$$Q - \bar{Q} = \sum_{i=1}^{r} p_i \left[\int_{M_i} f_i \, dE - \int_{\bar{M}_i} f_i \, dE \right].$$

Let $H = Q - \bar{Q}$. Since the estimates, $\bar{\theta}_i$, are random variables, H will be a random variable which is a function of the estimation functions, $\bar{\theta}_i$, as well as of the parameters, θ_i. The desired criterion for determining optimum estimates is then given by the following theorem:

THEOREM 2. *If* $E(\bar{\theta}_i - \theta_i)^4 = O(n^{-g})$, $g > 0$, *and if in some neighborhood of the point* $\bar{\theta}_i = \theta_i$, $(i = 1, \cdots, s)$ *the function* H *is continuous and possesses continuous derivatives of the first, second, and third order with respect to the* $\bar{\theta}_i$, *then*

$$E(H) = \frac{1}{2} \sum_{i=1}^{s} \sum_{j=1}^{s} H_{ij} E(\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j) + O(n^{-3/4g}),$$

where H_{ij} *denotes the partial derivative of* H *with respect to* $\bar{\theta}_j$ *and* $\bar{\theta}_i$ *at the point* $(\theta_1, \cdots, \theta_s)$.

The proof is similar to the type of proof used by Cramer [4] to obtain an expression for the variance of a function of central moments.

By means of Tchebycheff's inequality [4], page 182, it follows that

$$P[(\bar{\theta}_i - \theta_i)^4 \geq \epsilon^4] \leq \frac{E(\bar{\theta}_i - \theta_i)^4}{\epsilon^4}.$$

From the theorem assumptions, there exists a constant A such that

$$P[\bar{\theta}_i - \theta_i)^4 \geq \epsilon^4] < \frac{An^{-g}}{\epsilon^4}.$$

This is equivalent to

$$P[|\bar{\theta}_i - \theta_i| \geq \epsilon] < \frac{An^{-g}}{\epsilon^4}.$$

If E_1 denotes the set of points in sample space where $|\bar{\theta}_i - \theta_i| < \epsilon$, $(i = 1, \cdots, s)$, and E_2 denotes the complementary set, this inequality implies that

$$P[E_2] < \frac{sAn^{-g}}{\epsilon^4}. \tag{4}$$

The expected value of H may be written in the form

$$E(H) = \int_{E_1} H\, dP + \int_{E_2} H\, dP. \tag{5}$$

Consider the order of the second integral. From (4) and the fact that H is the difference of two probabilities, it follows that

$$\left| \int_{E_2} H\, dP \right| \leq \int_{E_2} dP = P[E_2] < \frac{sAn^{-\varrho}}{\epsilon^4}.$$

Consequently (5) becomes

$$E(H) = \int_{E_1} H\, dP + O(n^{-\varrho}). \tag{6}$$

Now consider the first integral. From the theorem assumptions, if ϵ is chosen sufficiently small, it follows that for any point in the set E_1, the function H can be expanded in the form

$$H = H(\theta) + \sum_1^s (\bar{\theta}_i - \theta_i)H_i(\theta) + \frac{1}{2}\sum_1^s \sum_1^s (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)H_{ij}(\theta) + R,$$

where θ denotes the point $(\theta_1, \cdots, \theta_s)$, where

$$R = \frac{1}{6}\sum_1^s \sum_1^s \sum_1^s (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)(\bar{\theta}_k - \theta_k)H_{ijk}(\theta'),$$

and where θ' is some point in E_1. Since $\bar{Q}$ reduces to Q when $\bar{\theta} = \theta$, $H(\theta) = 0$. Furthermore, since Q denotes the maximum probability of a correct classification, $H \geq 0$ for all $\bar{\theta}$; hence $H_i(\theta) = 0$ and $H_{ii}(\theta) \geq 0$ for all i. Thus, for any point in the set E_1,

$$H = \frac{1}{2}\sum_1^s \sum_1^s (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)H_{ij}(\theta) + R.$$

If this expression is substituted in (6), $E(H)$ will become

$$E(H) = \frac{1}{2}\sum_1^s \sum_1^s H_{ij}(\theta) \int_{E_1} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)\, dP + \int_{E_1} R\, dP + O(n^{-\varrho}). \tag{7}$$

Consider, first, the order of the remainder term. From the continuity assumption on H_{ijk}, it follows that H_{ijk} is bounded in E_1, say $| H_{ijk}(\theta') | < B$; hence

$$\left| \int_{E_1} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)(\bar{\theta}_k - \theta_k)H_{ijk}(\theta')\, dP \right| < B \int_{E_1} | (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)(\bar{\theta}_k - \theta_k) |\, dP.$$

By Schwarz's inequality,

$$\int_{E_1} | (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)(\bar{\theta}_k - \theta_k) |\, dP \leq \left[\int_{E_1} (\bar{\theta}_i - \theta_i)^2(\bar{\theta}_j - \theta_j)^2\, dP \int_{E_1} (\bar{\theta}_k - \theta_k)^2\, dP \right]^{\frac{1}{2}}.$$

Similarly,

$$\int_{E_1} (\bar{\theta}_i - \theta_i)^2(\bar{\theta}_j - \theta_j)^2\, dP \leq \left[\int_{E_1} (\bar{\theta}_i - \theta_i)^4\, dP \int_{E_1} (\bar{\theta}_j - \theta_j)^4\, dP\right]^{\frac{1}{2}},$$

$$\int_{E_1} (\bar{\theta}_k - \theta_k)^2\, dP \leq \left[\int_{E_1} (\bar{\theta}_k - \theta_k)^4\, dP \int_{E_1} dP\right]^{\frac{1}{2}} \leq \left[\int_{E_1} (\bar{\theta}_k - \theta_k)^4\, dP\right]^{\frac{1}{2}}.$$

Since

$$\int_{E_1} (\bar{\theta}_i - \theta_i)^4\, dP \leq \int_{E_1+E_2} (\bar{\theta}_i - \theta_i)^4\, dP = O(n^{-g}),$$

the preceding inequalities combine to give

$$\left|\int_{E_1} R\, dP\right| = O(n^{-3/4g}). \tag{8}$$

Now consider the first integral in (7). It may be written in the form

$$\int_{E_1} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)\, dP = E(\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j) - \int_{E_2} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)\, dP. \tag{9}$$

By Schwarz's inequality,

$$\left|\int_{E_2} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)\, dP\right| \leq \left[\int_{E_2} (\bar{\theta}_i - \theta_i)^2\, dP \int_{E_2} (\bar{\theta}_j - \theta_j)^2\, dP\right]^{\frac{1}{2}}.$$

Similarly,

$$\int_{E_2} (\bar{\theta}_i - \theta_i)^2\, dP \leq \left[\int_{E_2} (\bar{\theta}_i - \theta_i)^4\, dP \cdot P[E_2]\right]^{\frac{1}{2}}.$$

If these inequalities are combined and inequality (4) is employed, (9) will reduce to

$$\int_{E_1} (\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j)\, dP = E(\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j) + O(n^{-g}). \tag{10}$$

Finally, if (8) and (10) are employed in (7), it will reduce to the result stated in the theorem.

The order of the leading term in $E(H)$ depends upon the nature of the estimating functions, $\bar{\theta}_i$. In order to insure that this term will be the dominating term, and thus rule out pathological situations, only that class of estimating functions (estimators) will be considered for which this term will be of lower order than that of the remainder term. If the estimators are means or central moments, for example, then $g = 2$. For such estimators the order of the remainder term is $O(n^{-\frac{1}{2}})$, whereas the order of the leading term is not higher than $O(n^{-1})$.

A set of estimators will be called an optimum set if it maximizes the expected value of the probability of a correct classification, or, what is equivalent, if it minimizes $E(H)$. Since only large samples are being considered here, it is neces-

sary to define optimum in an asymptotic sense. Consider sets of estimators for which $E(H)$ is of order $O(n^{-q})$. For this class of estimators, a set will be called asymptotically optimum if it minimizes

$$\lim_{n\to\infty} n^q E(H).$$

Among asymptotically optimum sets of various orders, the set corresponding to the highest order would naturally be considered as the best asymptotic set. Now from Theorem 2, it readily follows that a set of estimators which minimizes

$$\sum_1^s \sum_1^s H_{ij} E(\bar{\theta}_i - \theta_i)(\bar{\theta}_j - \theta_j) \tag{11}$$

will be an asymptotically optimum set.

5. Maximum likelihood estimates. If the estimates $\bar{\theta}_i$ are unbiased and uncorrelated, (11) will reduce to

$$\sum_1^s H_{ii}\sigma_i^2 \tag{12}$$

where $\sigma_i^2 = E(\bar{\theta}_i - \theta_i)^2$ is a function of n as well as of the parameters. Since, from the discussion preceding (7), $H_{ii} \geq 0$, it follows that (12) will be a minimum when the σ_i^2 assume their minimum values. Now it is known [4], page 504, that under mild restrictions maximum likelihood estimates possess minimum asymptotic variances; hence for estimators of the type being considered which also satisfy the conditions in [4], the maximum likelihood estimates of the θ_i will yield an asymptotically optimum set of estimates for the classification problem.

REFERENCES

[1] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Roy. Soc. Phil. Trans.*, Vol. 231 (1933), pp. 289–337.

[2] B. L. Welch, "Note on discriminant functions," *Biometrika*, Vol. 31 (1939), pp. 218–220.

[3] A. Wald, "Contributions to the theory of statistical estimation and testing hypotheses," *Annals of Math. Stat.*, Vol. 10 (1939), pp. 299–304.

[4] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946, pp. 352–356.

NOTES

This section is devoted to brief research and expository articles on methology and other short items.

A GENERALIZATION OF WALD'S FUNDAMENTAL IDENTITY

By Gunnar Blom

University of Stockholm

1. Summary. The fundamental identity is generalized to the case of independent random variables with non-identical distributions. The conditions for the validity of the differentiation of the identity are discussed. The results given in [1], [2], and [3] are obtained as special cases.

2. A property of cumulative sums. Let $z_1, z_2, \cdots$ be an infinite sequence of independent random variables, $F_1(z), F_2(z), \cdots$ their distribution functions (d.f.) and $\varphi_1(t), \varphi_2(t), \cdots$ their moment-generating functions so that $\varphi_\nu(t) = E(e^{tz_\nu})$. a_N and b_N are given constants ($a_N > b_N$, $N = 1, 2, \cdots$). n is defined as the smallest integer N for which $Z_N = z_1 + \cdots + z_N$ is $\geqq a_N$ or $\leqq b_N$.

We first give two lemmas.

Lemma 1. *If two positive quantities δ and ϵ can be found such that one at least of the following conditions a) and b) are satisfied*

$$a)\ P(z_\nu > \delta) > \epsilon \text{ for all } \nu \text{ and } \limsup_{N\to\infty} a_N < \infty$$

$$b)\ P(z_\nu < -\delta) > \epsilon \text{ for all } \nu \text{ and } \liminf_{N\to\infty} b_N > -\infty,$$

then for any $k \geqq 0$

$$\lim_{N\to\infty} N^k P(n > N) = 0. \tag{1}$$

An inspection of the proof of (4) in [4] shows that this formula holds when the conditions of the lemma are satisfied. The lemma follows.

Lemma 1 can be generalized as follows.

Lemma 2. *If two positive quantities δ and ϵ and a sequence $c_1, c_2, \cdots$ can be found such that one at least of the following conditions a) and b) are satisfied*

$$a)\ P(z_\nu + c_\nu > \delta) > \epsilon \text{ for all } \nu, \quad \limsup_{N\to\infty} a_N < \infty, \quad \limsup_{N\to\infty} \sum_1^N c_\nu < \infty,$$

$$b)\ P(z_\nu + c_\nu < -\delta) > \epsilon \text{ for all } \nu,$$

$$\liminf_{N\to\infty} b_N > -\infty, \quad \liminf_{N\to\infty} \sum_1^N c_\nu > -\infty,$$

then (1) *is true.*

PROOF: In case a) we put $z_\nu' = z_\nu + c_\nu$, $Z_N' = \Sigma\, z_\nu'$ and $a_N' = a_N + \Sigma_1^N c_\nu$. The inequality $Z_N \geqq a_N$ then becomes $Z_N' \geqq a_N'$. As $P(z_\nu' > \delta) > \epsilon$ and $\limsup_{N\to\infty} a_N' < \infty$, Lemma 1 can be applied to the sequence $z_1', z_2', \cdots$, and thus (1) is true. When conditions b) are satisfied, the proof is analogous.

3. The generalized fundamental identity. In this section we shall consider sequences of random variables of the type defined in Lemma 2. We shall prove two theorems the first of which is valid for complex values of t and the second only for real values of t.

THEOREM 1. *Assuming that*

1°. *one at least of conditions a) and b) of Lemma 2 is satisfied;*

2°. $b \leqq b_N < a_N \leqq a$, *where a and b are finite;*

3°. *for some complex (or real) value of t,* $\varphi_\nu(t)$ *exists for all* ν *and is* $\neq 0$ *and*

$$\liminf_{N\to\infty} |\, \varphi_1(t)\cdots\varphi_N(t)\,| > 0,$$

then

$$E[e^{tZ_n}(\varphi_1(t)\cdots\varphi_n(t))^{-1}] = 1. \tag{2}$$

PROOF. Let W_m denote the set of all sequences $z_1 \cdots z_N$ in the N-dimensional Euclidean space Ω_N for which $n = m$ $(m \leqq N)$, W_m' the projection of W_m on Ω_m and $W_{n>N}$ all sequences for which $n > N$. We have identically

$$\left[\sum_{m=1}^{N}\int_{W_m} + \int_{W_{n>N}}\right] e^{tZ_N}\, dF_1 \cdots dF_N = \int_{\Omega_N} e^{tZ_N}\, dF_1 \cdots dF_N = \varphi_1(t)\cdots\varphi_N(t).$$

Dividing by the right member and cancelling common factors we obtain

$$\sum_{m=1}^{N} (\varphi_1\cdots\varphi_m)^{-1}\int_{W_m'} e^{tZ_m}\, dF_1\cdots dF_m + (\varphi_1\cdots\varphi_N)^{-1}\int_{W_{n>N}} e^{tZ_N}\, dF_1\cdots dF_N = 1. \tag{3}$$

When $N \to \infty$ the first sum tends to the left member of (2). We thus have to investigate the last term in (3) which we denote by R_N. We can write

$$R_N = (\varphi_1\cdots\varphi_N)^{-1}\int_{W_{n>N}} e^{tZ_N}\, dF_1\cdots dF_N = (\varphi_1\cdots\varphi_N)^{-1} P(n > N) E_{n>N} e^{tZ_N}. \tag{4}$$

It follows from Lemma 2 that $P(n > N) \to 0$. As $b < Z_N < a$ by 2° we conclude that $R_N \to 0$. This proves the theorem.

THEOREM 2. *If, for some real value of t,* $\varphi_\nu(t)$ *exists for all* ν *and if quantities* c_ν, $\epsilon > 0$ *and* $\delta > 0$ *can be found such that at least one of the following conditions a) and b) are satisfied for all* ν

a) $\limsup_{N\to\infty} a_N < \infty$, $\limsup_{N\to\infty} \sum_1^N c_\nu < \infty$ *and*

$$A_\nu(t,\delta) = \frac{1}{\varphi_\nu(t)}\int_{\delta - c_\nu}^{\infty} e^{tz}\, dF_\nu(z) > \epsilon, \qquad (\nu = 1, 2, \cdots), \tag{5a}$$

b) $\liminf_{N\to\infty} b_N > -\infty$, $\liminf_{N\to\infty} \sum_1^N c_\nu > -\infty$ *and*

$$B_\nu(t, \delta) = \frac{1}{\varphi_\nu(t)} \int_{-\infty}^{-\delta - c_\nu} e^{tz}\, dF_\nu(z) > \epsilon, \qquad (\nu = 1, 2, \cdots), \tag{5b}$$

then (2) *holds.*

The conditions of the theorem become more attractive if the theorem is limited to the somewhat less general cases mentioned in the Corollary below. The above formulation has been chosen mainly because of an important application to identical variables in Sec. 6.

PROOF. The theorem is proved if we can show that R_N in (4) tends to zero when $N \to \infty$. For that purpose we use the transformation (cf [5] and [3])

$$G_\nu(z; t) = \frac{1}{\varphi_\nu(t)} \int_{-\infty}^{z} e^{tz}\, dF_\nu(z), \qquad (\nu = 1, 2, \cdots). \tag{6}$$

$G_\nu(z; t)$ is obviously a d.f. for every real t (for which $\varphi_\nu(t)$ exists). When (5a) holds,

$$P[z_\nu + c_\nu > \delta \mid G_\nu(z; t)] = A(t, \delta).$$

Here the expression in the left member denotes the probability that $z_\nu + c_\nu > \delta$, when G_ν is the d.f. of z_ν.

Consequently, when conditions a) are fulfilled, a sequence of random variables with the d.f:s $G_1(z; t)$, $G_2(z; t)$, $\cdots$ or, with one notation, $G(t)$ satisfies the conditions a) of Lemma 2. It follows that

$$\lim_{N\to\infty} P(n > N \mid G(t)) = 0.$$

Introducing $G_\nu(z; t)$ in R_N we find

$$R_N = \int_{W_n > N} dG_1 \cdots dG_N = P(n > N \mid G(t)).$$

Consequently $R_N \to 0$. When conditions b) are fulfilled, the proof is analogous.

COROLLARY TO THEOREM 2. *If* $1°$ $\varphi_\nu(t)e^{tc_\nu} \leqq H(t) < \infty$, $2°$ t *is positive and conditions a) of Lemma* 2 *hold or* t *is negative and conditions b) of Lemma* 2 *hold, then the generalized fundamental identity is true.*

For, in the first case

$$A_\nu(t, \delta) \geqq \frac{e^{t(\delta - c_\nu)}}{\varphi_\nu(t)} \int_{\delta - c_\nu}^{\infty} dF_\nu \geqq \frac{\epsilon e^{t\delta}}{H(t)} = \epsilon_1(t)$$

so that (5a) is satisfied, and similarly when t is negative.

The following special case deserves particular attention as it covers most cases occurring in practice and the conditions become very simple: *If a sequence of random variables satisfies conditions a) and b) of Lemma* 1 *simultaneously, a sufficient condition for the validity of* (2) *for some given real value of* t *is that the sequence* $\varphi_\nu(t)$ *is bounded.*

4. Application to Poisson variables. As an application of (2) we consider a sequence of Poisson variables with the parameters λm_ν, where λ is a positive quantity and m_ν are positive integers. From the well-known formula

$$\varphi_\nu(t) = e^{\lambda m_\nu (e^t - 1)}$$

we easily conclude that the conditions of Theorem 1 are valid if $R(e^t) \geqq 1$. (With $\delta < 1$ in (5a) we find that (2) holds even for negative t.) If, in particular, we choose t so that $e^t = 1 + \dfrac{2\pi i k}{\lambda} = c_k$, we have the simple formula

$$E(c_k^{Z_n}) = 1, \qquad (k = 1, 2, \cdots).$$

5. Differentiation of the generalized fundamental identity. In this section t is assumed to be real. We denote the kth derivative of $\varphi_\nu(t)$ by $\varphi_\nu^{(k)}(t)$. We shall prove the following theorem which corresponds to Theorems 1 and 2.

THEOREM 3. *If for all t in a closed interval I the conditions stated in Theorems 1 or 2 are satisfied and if, in addition, the functions* $\left|\dfrac{\varphi_\nu^{(k)}(t)}{\varphi_\nu(t)}\right|$ *are uniformly bounded with respect to both ν and t (in I) for $k = 1, 2, \cdots r$, then the generalized fundamental identity may be differentiated r times with respect to t for any t in the interior of I.*

We use a method of proof which is similar to that used in [2]. We first show that the sum in (3) may be differentiated r times under the integral signs and secondly that the rth derivative of R_N tends to zero uniformly in t when $N \to \infty$.

The rth derivative of the general term of the series in (3) consists of a finite number of terms of the form

$$J_m(t) = (\varphi_1 \cdots \varphi_m)^{-1} H_\mu \int_{W_m'} Z_m^\lambda e^{tZ_m}\, dF_1 \cdots dF_m \qquad (\mu \leqq \lambda; \mu, \lambda = 1, 2, \cdots r),$$

and the rth derivative of R_N in (4) consists of a finite number (which does not depend on N) of similar expressions with N substituted for m and $W_{n>N}$ for W_m'. H_μ is a sum of m^μ and N^μ terms respectively which is symmetric in ν. The terms are functions of $\dfrac{\varphi_\nu^{(k)}(t)}{\varphi_\nu(t)}$ $(k \leqq \lambda; \nu = 1, 2, \cdots m)$ and are thus majorated by the same constant C.

Further, we can always find a positive quantity t_0 such that for all t in I

$$|\, Z_m^\lambda e^{tZ_m} \,| \leqq e^{t_0 |Z_m|} \leqq (e^{t_0 Z_m} + e^{-t_0 Z_m}).$$

Hence

$$|\, J_m(t) \,| \leqq (\varphi_1 \cdots \varphi_m)^{-1} C m^\mu \int_{W_m'} (e^{t_0 Z_m} + e^{-t_0 Z_m})\, dF_1 \cdots dF_m. \tag{7}$$

The rest of the proof is divided into two parts corresponding to the conditions of Theorem 1 and those of Theorem 2.

When the conditions of Theorem 2 are fulfilled we make the transformation (6) in (7) with $t = t_0$ and $t = -t_0$. Then

$$| J_m(t) | \leqq Cm^{\mu}[P(n = m \mid G(t_0)) + P(n = m \mid G(-t_0))] \leqq 2Cm^{\mu} < \infty.$$

This justifies the differentiation of the series in (3).

Substituting N for m and $n > N$ for $n = m$ in the above expression we further have

$$| J_N(t) | \leqq CN^{\mu}[P(n > N \mid G(t_0)) + P(n > N \mid G(-t_0))],$$

and conclude from Lemma 2 with $k = \mu$ in (1) that $J_N(t)$ tends to zero uniformly in t. It follows that the rth derivative of R_N also tends to zero uniformly in t.

In the second part of the proof we assume the conditions of Theorem 1 to be satisfied. We then write (7) in the following form

$$| J_m(t) | \leqq C(\varphi_1 \cdots \varphi_m)^{-1} m^{\mu} P(n = m) E_{n=m}(e^{t_0 Z_m} + e^{-t_0 Z_m}), \tag{8}$$

where $E_{n=m}$ signifies the conditional expectation when it is known that $n = m$. From the definition of n it follows that, when $n = m$, we have $b_{m-1} < Z_{m-1} < a_{m-1}$ and $Z_m \geqq a_m$ or $\leqq b_m$. Hence

$$E_{n=m}(e^{t_0 Z_m}) \leqq E_{n=m}(e^{t_0 Z_m} \mid Z_m \geqq a_m) = E_{n=m}[e^{t_0(Z_{m-1}+z_m)} \mid Z_{m-1} + z_m \geqq a_m]$$
$$\leqq e^{t_0 a_m - 1} E[e^{t_0 z_m} \mid z_m > a_m - b_{m-1}] < \infty.$$

The second exponential can be treated in a similar way. Thus $J_m(t)$ is majorated by a finite expression.

Finally, we substitute N for m and $n > N$ for $n = m$ in (8). I being a closed interval it follows from condition 3° in Theorem 1 that we can find a constant C such that

$$| J_N(t) | \leqq CN^{\mu} P(n > N) E_{n>N}(e^{t_0 Z_N} + e^{-t_0 Z_N}).$$

From the definition of n and condition 2° in Theorem 1 we have $b < Z_N < a$. An application of Lemma 2 then shows that $J_N(t)$ tends to zero uniformly in t. This proves the theorem.

COROLLARY TO THEOREM 3. *When the conditions stated in Corollary of Theorem* 2 *are fulfilled for all* t *in the closed interval* I, *Theorem* 3 *is true.*

This is obvious.

6. The fundamental identity for identically distributed variables. In the special case of identically distributed variables for which $P(z = 0) < 1$ and $0 < \varphi(t) < \infty$ we infer from Theorem 1 that the fundamental identity

$$E[e^{tZ_n}(\varphi(t))^{-n}] = 1 \tag{9}$$

holds if t is complex and $| \varphi(t) | \geqq 1$. This is the case discussed in [1].

Further, when $P(z = 0) < 1$, the integrals $\int_a^{\infty} e^{tz} dF$ and $\int_{-\infty}^{\beta} e^{tz} dF$ cannot both

be zero for every $\alpha > 0$ and $\beta < 0$, and thus we infer from Theorem 2 that the fundamental identity holds for all real t (if the limits a_N and b_N are chosen in accordance with the conditions of this theorem). This proposition is somewhat more general than that proved in [3] by a similar method.

It also follows from the last remark and Theorem 3 that, when $P(z = 0) < 1$, (9) can be differentiated any number of times for any real t. This proposition contains the results in [2] and [3] as special cases.

7. A generalization. We finally remark that the assumption made in Theorem 3 that the expressions containing derivatives of $\varphi_\nu(t)$ are uniformly bounded is unnecessarily restrictive. For example, it seems possible to prove that the first derivative of (2) may be obtained by differentiation under the expectation sign if the series (cf. Corollary 1 to Theorem 7.4. in [6])

$$\sum_{m=1}^{\infty} P(n = m) \sum_{\nu=1}^{m} \frac{\varphi_\nu'(t)}{\varphi_\nu(t)}$$

is uniformly convergent with respect to t.

REFERENCES

[1] A. Wald, "On cumulative sums of random variables," *Annals of Math. Stat.*, Vol. 15 (1944), p. 285.

[2] A. Wald, "Differentiation under the expectation sign in the fundamental identity of sequential analysis," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 493–497.

[3] G. E. Albert, "A note on the fundamental identity of sequential analysis," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 593–596 and Vol. 19 (1948), pp. 426–427.

[4] C. Stein, "A note on cumulative sums," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 498–499.

[5] H. Cramér, Sur un nouveau theorème-limite de la théorie des probabilités, *Actualités, scientifiques et industrielles*, no. 736, Hermann et Cie., 1938, p. 5.

[6] J. Wolfowitz, "The efficiency of sequential estimates and Wald's equation for sequential processes," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 228–229.

SPREAD OF MINIMA OF LARGE SAMPLES

By Brockway McMillan

Bell Telephone Laboratories, Murray Hill, N. J.

1. Theorems. Let x have the continuous cumulative distribution function $F(x)$. Let $(x_1, \cdots, x_N)$ be a sample of N independent values of x and $y = \inf (x_1, \cdots, x_N)$. Then y is a random variable with the cumulative distribution function

$$G_N(y) = 1 - (1 - F(y))^N. \tag{1}$$

Let K values of the new variable y be drawn, $(y_1, \cdots, y_K)$ and let the spread

$$w = \sup (y_1, \cdots, y_K) - \inf (y_1, \cdots, y_K).$$

Fixing K, we consider the cumulative distribution function of w, $P_N(w)$, as $N \to \infty$. That is, we have K large samples of x and wish to examine the spread among their minima. It is evident intuitively that if $F(x) = 0$ for some finite x, these minima are bounded from below and will cluster near the vanishing point of $F(x)$, making $w \to 0$ statistically as $N \to \infty$. Our theorems also show that even when $y \to -\infty$ statistically, i.e., when $F(x) = 0$ for no finite x, the spread $w \to 0$ statistically if the tail of $F(x)$ is sufficiently small (e.g. Gaussian). On the other hand, if $F(x) = 0(e^{kx})$ as $x \to -\infty$, the distribution $P_N(w)$ does not peak as $N \to \infty$, while for larger tails (e.g. algebraic) $w \to +\infty$ statistically. Two simple theorems are

I. *If*

$$\lim_{x \to -\infty} \frac{F(x)}{F(x+s)} = 1,$$

then

$$\lim_{N \to \infty} P_N(s) = 0.$$

II. *Let* $s > 0$. *If*

$$F(x_0) = 0 \text{ for some } x_0 > -\infty, \text{ or if}$$

$$\lim_{x \to -\infty} \frac{F(x)}{F(x+s)} = 0,$$

then

$$\lim_{N \to \infty} P_N(s) = 1.$$

Theorem I is directly applicable to distributions with algebraic tails, theorem II to Gaussian tails. We prove them both as corollaries of the more general results:

III. *If*

$$\liminf_{x \to -\infty} \frac{F(x)}{F(x+s)} = l$$

then

$$\limsup_{N \to \infty} P_N(s) \leq (1-l)^{K-1}.$$

IV. *Let* $s > 0$. If

$$F(x) = 0 \text{ for no finite } x \text{ and}$$

$$\limsup_{x \to -\infty} \frac{F(x)}{F(x+s)} = L,$$

then

$$\liminf_{N \to \infty} P_N(s) \geq [e^{-\alpha L} - e^{-\alpha}]^K$$

for any $\alpha > 0$.

Theorems III and IV together show that an exponential tail ($F(x) = 0(e^{kx})$) leads to a $P_N(w)$ which, asymptotically, is bounded away from 0 for any $w > 0$ and bounded away from 1 for w sufficiently small.

2. Proofs. Explicitly, for any $s \geq 0$,

$$P_N(s) = K \int_{-\infty}^{\infty} [G_N(x+s) - G_N(x)]^{K-1}\, dG_N(x+s). \tag{2}$$

Turning now to III: given $s > 0$, choose $x_1 = x_1(\epsilon)$ so that (i) $F(x_1) \neq 0$, and (ii), $x \leq x_1$ implies

$$\frac{F(x)}{F(x+s)} \geq l - \epsilon. \tag{3}$$

We then rewrite (2) as

$$P_N(s) = \int_{-\infty}^{x_1} \left[1 - \frac{G_N(x)}{G_N(x+s)}\right]^{K-1} dG_N(x+s)^K + \int_{x_1}^{\infty}. \tag{4}$$

Treating $G_N(x+s)^K$ as the independent variable, the first integral may be evaluated by the mean value theorem in the form

$$\left[1 - \frac{G_N(x_2)}{G_N(x_2+s)}\right]^{K-1} \int_{-\infty}^{x_1} dG_N(x+s)^K \leq \left[1 - \frac{G_N(x_N)}{G_N(x_N+s)}\right]^{K-1} \tag{5}$$

with an appropriate $x_2 = x_2(N)$, $-\infty \leq x_2 \leq x_1$.

Using the form (2) of the integrand in the second term of (4), we may bound the latter by

$$K \int_{x_1}^{\infty} dG_N(x+s) \leq K[1 - G_N(x_1+s)], \tag{6}$$

since

$$G_N(x+s) - G_N(x) \leq 1.$$

Now, by factoring (1),

$$\frac{G_N(x)}{G_N(x+s)} = \frac{F(x)}{F(x+s)} \frac{1 + Q + \cdots + Q^{N-1}}{1 + Q_s + \cdots + Q_s^{N-1}} \geq \frac{F(x)}{F(x+s)} \tag{7}$$

where $Q = 1 - F(x)$, $Q_s = 1 - F(x+s) \leq Q$. Combining (3), (4), (5), (6), and (7),

$$P_N(s) \leq [1 - l + \epsilon]^{K-1} + K[1 - G_N(x_1+s)].$$

Since $F(x_1+s) \geq F(x_1) > 0$, we have

$$\lim_{N\to\infty} G_N(x_1+s) = 1.$$

Hence,

$$\limsup_{N\to\infty} P_N(s) \leq [1 - l + \epsilon]^{K-1}$$

and III follows by letting $\epsilon \to 0$. Then I follows immediately with $l = 1$, when we note that $P_N(s) \geq 0$.

To prove IV, choose any $\alpha > 0$. By hypothesis, for sufficiently large N we may always find $x_N = x_N(\alpha)$ such that

$$F(x_N) = \frac{\alpha L}{N}. \tag{8}$$

By hypothesis, and the monotonicity of $F(x)$, $x_N \to -\infty$ as $N \to \infty$. For any $\epsilon > 0$, therefore, we can find $N_0 = N_0(\alpha, \epsilon)$ such that $N \geq N_0$ implies

$$\frac{F(x_N)}{F(x_N + s)} \leq \frac{L}{1 - \epsilon} \tag{9}$$

or $F(x_N + s) \geq \frac{\alpha}{N}(1 - \epsilon)$. Directly from (2), since $s > 0$,

$$P_N(s) \geq K \int_{x_N - s}^{x_N} [G_N(x + s) - G_N(x)]^{K-1}\, dG_N(x + s)$$

$$\geq K \int_{x_N - s}^{x_N} [G_N(x + s) - G_N(x_N)]^{K-1}\, dG_N(x + s).$$

But this last integral is of the form

$$\int K(U - G)^{K-1}\, dU = (U - G)^K,$$

whence

$$P_N(s) \geq [G_N(x_N + s) - G_N(x_N)]^K,$$

or

$$P_N(s) \geq [(1 - F(x_N))^N - (1 - F(x_N + s))^N]^K. \tag{10}$$

By (8) and (9), therefore

$$P_N(s) \geq \left[\left(1 - \frac{\alpha L}{N}\right)^N - \left(1 - \frac{\alpha(1 - \epsilon)}{N}\right)^N\right]^K$$

Since this holds for all $N \geq N_0(\alpha, \epsilon)$,

$$\liminf_{N \to \infty} P_N(s) \geq [e^{-\alpha L} - e^{-\alpha(1-\epsilon)}]^K$$

This last, in turn, now holds for any $\epsilon > 0$, hence

$$\liminf_{N \to \infty} P_N(s) \geq [e^{-\alpha L} - e^{-\alpha}]^K.$$

This now holds for any $\alpha > 0$. Maximizing on α yields a sharper bound than the result of IV. The applicable part of II follows, when $L = 0$, by letting $\alpha \to \infty$. That the conclusion of II holds when $F(x_0) = 0$ for some finite x_0 follows from (10) with x_N replaced by some x_1 such that $F(x_1) = 0$, $F(x_1 + s) > 0$.

ON THE CONVERGENCE OF THE CLASSICAL ITERATIVE METHOD OF SOLVING LINEAR SIMULTANEOUS EQUATIONS[1]

By Edgar Reich

Massachusetts Institute of Technology

The classical iterative method, or Seidel method, is a scheme for solving the system of linear algebraic equations

$$\sum_{j=1}^{n} A_{ij}x_j = b_i, \qquad (i = 1, 2, \cdots, n),$$

by successive approximation, as follows:
If $x^{(\nu)} = (x_1^{(\nu)}, x_2^{(\nu)}, \cdots, x_n^{(\nu)})$ is the νth approximation of the solution, the $(\nu + 1)$st approximation, $x^{(\nu+1)} = (x_1^{(\nu+1)}, x_2^{(\nu+1)}, \cdots, x_n^{(\nu+1)})$, is obtained from the relations

$$\begin{cases} A_{11}x_1^{(\nu+1)} + A_{12}x_2^{(\nu)} + A_{13}x_3^{(\nu)} + \cdots + A_{1n}x_n^{(\nu)} = b_1, \\ A_{21}x_1^{(\nu+1)} + A_{22}x_2^{(\nu+1)} + A_{23}x_3^{(\nu)} + \cdots + A_{2n}x_n^{(\nu)} = b_2, \\ A_{31}x_1^{(\nu+1)} + A_{32}x_2^{(\nu+1)} + A_{33}x_3^{(\nu+1)} + \cdots + A_{3n}x_n^{(\nu)} = b_3, \\ \cdots \\ A_{n1}x_1^{(\nu+1)} + A_{n2}x_2^{(\nu+1)} + A_{n3}x_3^{(\nu+1)} + \cdots + A_{nn}x_n^{(\nu+1)} = b_n, \end{cases}$$

$x_1^{(\nu+1)}$ being obtained from the first equation, then $x_2^{(\nu+1)}$ from the second, and so on.

The given system can be written in matrix notation as $Ax = b$ where A is a non-singular square matrix of order n, and x and b are column vectors of order n. Let us define square matrices A_1 and A_2 as follows:

$$(A_1)_{ij} = \begin{cases} A_{ij} \text{ if } i \geq j \\ 0 \text{ if } i < j \end{cases},$$

$$(A_2)_{ij} = \begin{cases} A_{ij} \text{ if } i < j \\ 0 \text{ if } i \geq j \end{cases},$$

(Note that $A_1 + A_2 = A$.).

With this notation the Seidel method can be written as the matric difference equation

$$A_1x^{(\nu+1)} + A_2x^{(\nu)} = b.$$

Now various writers, among them C. E. Berry in this journal, (See list of refer-

[1] Work done under Office of Naval Research Contract N5ori60.

ences at end of this paper.) have shown that a necessary and sufficient condition for convergence, i.e., a necessary and sufficient condition for

$$\lim_{\nu\to\infty} (x_i^{(\nu)} - x_i) = 0, \qquad (i = 1, 2, \cdots, n),$$

is that

(1) A_1 has an inverse; that is $A_{ii} \neq 0$ for any i.

(2) The characteristic roots of $(A_1^{-1}A_2)$ all have an absolute value smaller than unity.

It would be advantageous to rephrase the above condition, if possible, in terms of simpler requirements on A. As a step in this direction the following theorem is offered:

THEOREM. *If A is a real, symmetric nth-order matrix with all terms on its main diagonal positive, then a necessary and sufficient condition for all the n characteristic roots of $(A_1^{-1}A_2)$ to be smaller than unity in magnitude is that A is positive definite.*

PROOF. Let z_j be a characteristic vector of $(A_1^{-1}A_2)$ corresponding to the characteristic root μ_j. Then

$$(A_1^{-1}A_2)\, z_j = \mu_j z_j . \tag{1}$$

Premultiplying by $\bar{z}_i' A_1$, where the apostrophe and bar denote transposition and conjugation respectively:

$$\bar{z}_i' A_2 z_j = \mu_j \bar{z}_i' A_1 z_j . \tag{2}$$

Consider the bilinear form $\bar{z}_i' A z_j$.
We have

$$\bar{z}_i' A z_j = \bar{z}_i' A_1 z_j + \bar{z}_i' A_2 z_j = (1 + \mu_j)\, \bar{z}_i' A_1 z_j . \tag{3}$$

Interchanging i and j:

$$\bar{z}_j' A z_i = (1 + \mu_i) \bar{z}_j' A_1 z_i . \tag{4}$$

Taking the conjugate:

$$z_j' A \bar{z}_i = \bar{z}_i' A z_j = (1 + \bar{\mu}_i) z_j' A_1 \bar{z}_i = (1 + \bar{\mu}_i) \bar{z}_i' A_1' z_j . \tag{5}$$

Let D be the diagonal matrix with elements

$$D_{ij} = A_{ij}\delta_{ij} . \tag{6}$$

This makes $A_1' = D + A_2$.
Substituting this in (5):

$$\bar{z}_i' A z_j = (1 + \bar{\mu}_i)\, (\bar{z}_i' D z_j + \bar{z}_i' A_2 z_j) = (1 + \bar{\mu}_i) \bar{z}_i' D z_j + (1 + \bar{\mu}_i) \mu_j \bar{z}_i' A_1 z_j . \tag{7}$$

Eliminating $\bar{z}_i' A_1 z_j$ between relations (3) and (7) we obtain

$$(1 - \bar{\mu}_i \mu_j) \bar{z}_i' A z_j = (1 + \bar{\mu}_i)\, (1 + \mu_j) \bar{z}_i' D z_j . \tag{8}$$

To obtain the necessary condition we use the fact that we must have $|\mu_i| < 1$, and can therefore rewrite (8) as

$$\bar{z}_i' A z_j = \frac{(1+\bar{\mu}_i)(1+\mu_j)}{1-\bar{\mu}_i\mu_j}\,\bar{z}_i' D z_j = \sum_{k=0}^{\infty} (1+\bar{\mu}_i)\bar{\mu}_i^k(1+\mu_j)\mu_j^k \bar{z}_i' D z_j. \tag{9}$$

If $x = \sum_{i=1}^{m} c_i z_i$ is any linear combination of the $m \leq n$ independent characteristic vectors of $(A_1^{-1}A_2)$ then

$$\begin{aligned}\bar{x}' A x &= \left(\sum_{i=1}^{m} \bar{c}_i \bar{z}_i'\right) A \left(\sum_{i=1}^{m} c_i z_i\right) = \sum_{i,j=1}^{m} \bar{c}_i c_j \bar{z}_i' A z_j \\ &= \sum_{i,j=1}^{m} \bar{c}_i c_j \sum_{k=0}^{\infty} (1+\bar{\mu}_i)\bar{\mu}_i^k(1+\mu_j)\mu_j^k \bar{z}_i' D z_j,\end{aligned} \tag{10}$$

or

$$\bar{x}' A x = \sum_{k=0}^{\infty} \bar{y}_k' D y_k$$

where

$$y_k = \sum_{i=1}^{m} c_i(1+\mu_i)\mu_i^k z_i.$$

Since by hypothesis $A_{ii} > 0$, D is evidently positive definite, and therefore

$$\bar{x}' A x > 0. \tag{11}$$

In case the characteristic roots μ_i, $(i = 1, 2, \cdots n)$, are all distinct there will be n independent z_i assured, and in that case (11) implies that A is positive definite. Consider, on the other hand, the case where the μ_i are not all distinct. Note that (a) the definiteness properties of a matrix are not changed by sufficiently small alterations in the elements; (b) the μ's depend continuously on the elements of A; (c) the discriminant of (1) is a polynomial in the A_{ij} that does not vanish identically.[2] It follows that A must be positive definite even in the case of repeated roots because an arbitrarily small change in A will separate any multiple μ's, still keeping them smaller than unity in magnitude, and not changing the definiteness properties of A.

This completes the proof that the condition given in the statement of the theorem is necessary. Now to prove sufficiency:

Setting $i = j$ in relation (8) we obtain

$$(1 - |\mu_i|^2)\bar{z}_i' A z_i = |1+\mu_i|^2 \bar{z}_i' D z_i \tag{12}$$

Since both A and D are positive definite

$$\bar{z}_i' A z_i > 0 \text{ and } \bar{z}_i' D z_i > 0. \tag{13}$$

[2] The fact that the discriminant is not identically zero follows from easily constructible counter-examples.

Moreover, we cannot have $\mu_i = -1$ because that would mean by (3) that

$$0 = \bar{z}_i' A_1 z_i + \bar{z}_i' A_2 z_i = \bar{z}_i' A z_i .$$

Relation (12) thus implies

$$1 - |\mu_i|^2 > 0 \tag{14}$$

i.e. $|\mu_i| < 1$ as was to be proved.

The part of the theorem giving the sufficient condition was already obtained by L. Seidel [1] and G. Temple in a somewhat more indirect fashion.

REFERENCES

[1] L. Seidel, "Über ein Verfahren die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineare Gleichungen überhaupt, durch successive Annäherung aufzulösen," *Abhandlungen der Mathematisch-Physikalischen Classe der Königlich Bayerischen Akademie der Wissenschaften*, Vol. 11 (1874), pp. 81–108.

[2] C. E. Berry, "A criterion of convergence for the classical iterative method of solving linear simultaneous equations," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 398–400.

[3] L. Cesari, "Sulla risoluzione dei sistemi di equazioni lineari per approssimazioni successive," *Rassegna delle Poste, dei Telegrafi e dei Telefoni*, Anno 9 (1931).

[4] L. Cesari, "Sulla risoluzione dei sistemi di equazioni lineari per approssimazioni successive," *Reale Accademia Nazionale dei Lincei, Serie 6, Classe di Szienze fisiche, matematiche e naturali, Rendiconti*, Vol. 25 (1937), pp. 422–428.

[5] J. Morris, *The Escalator Method in Engincering Vibration Problems*, Chapman and Hall Ltd., London 1947, pp. 63–70.

[6] R. J. Schmidt, "On the numerical solution of linear simultaneous equations by an iterative method," *Phil. Mag.*, Ser. 7, Vol. 32 (1941), pp. 369–383.

SOME RECURRENCE FORMULAE IN THE INCOMPLETE BETA FUNCTION RATIO

By T. A. Bancroft

Alabama Polytechnic Institute

1. Introduction. It is well known that the incomplete beta function ratio, defined by

$$I_x(p, q) = \frac{B_x(p, q)}{B(p, q)}, \tag{1}$$

where

$$B_x(p, q) = \int_0^x x^{p-1}(1 - x)^{q-1}\, dx, \tag{2}$$

and

$$B(p, q) = B_1(p, q), \tag{3}$$

is of importance in probability distribution theory, and, hence, also in obtaining exact probability values in making tests of statistical hypotheses. In constructing certain extensions [1] of Karl Pearson's "Tables of the Incomplete Beta-Function" [2], the recurrence formulae contained in the following sections were derived.

2. Derivation of formulae. The incomplete beta function, $B_x(p, q)$ may be considered as a special case of the hypergeometric series, $F(a, b, c, x)$, thus

$$B_x(p, q) = \frac{x^p}{p} F(p, 1 - q, p + 1, x). \tag{4}$$

The series converges for $|x| \leq 1$, if and only if $a + b < c$. By setting $a = p$, $b = 1 - q$, and $c = p + 1$, as in (4), all conditions are satisfied, if we also take $q > 0$.

Recurrence formulae for $F(a, b, c, x)$, e. g., in the work of Magnus and Oberhettinger [3], may now be directly converted for use with $B_x(p, q)$ or $I_x(p, q)$. In particular, using the three identities on page 9 of [3], with x replacing z, we have

$$cF(a, b, c, x) + (b - c)F(a + 1, b, c + 1, x) - b(1 - x)F(a + 1, b + 1, c + 1, x) = 0, \tag{5}$$

$$c(c - ax - b)F(a, b, c, x) - c(c - b)F(a, b - 1, c, x) + abx(1 - x)F(a + 1, b + 1, c + 1, x) = 0, \tag{6}$$

$$cF(a, b, c, x) - cF(a, b + 1, c, x) + axF(a + 1, b + 1, c + 1, x) = 0, \tag{7}$$

with $a = p$, $b = 1 - q$, and $c = p + 1$, we obtain in turn

$$xI_x(p, q) - I_x(p + 1, q) + (1 - x)I_x(p + 1, q - 1) = 0 \tag{8}$$

$$(p + q - px)I_x(p, q) - qI_x(p, q + 1) - p(1 - x)I_x(p + 1, q - 1) = 0 \tag{9}$$

$$qI_x(p, q + 1) + pI_x(p + 1, q) - (p + q)I_x(p, q) = 0. \tag{10}$$

Formula (8) is the basic recurrence formula used in the construction of Karl Pearson's [2] tables. Formula (10) was obtained, incidentally, by the author [4] in a different connection and manner.

Formulae (8), (9), and (10) may now be combined to give other useful formulae, e. g.,

$$qI_x(p + 1, q + 1) + (\overline{p + q}x - q)I_x(p + 1, q) - (p + q)xI_x(p, q) = 0, \tag{11}$$

$$pI_x(p + 1, q + 1) + (q - \overline{p + q}x)I_x(p, q + 1) - (p + q)(1 - x)I_x(p, q) = 0, \tag{12}$$

$$(13)\quad (p+q-1)xI_x(p-1,q) - (\overline{p+q-1}x+p)I_x(p,q) + pI_x(p+1,q) = 0,$$

$$(14)\quad (p+q)(1-x)I_x(p+1,q-1) - \{(p+q)(1-x)+q\}I_x(p+1,q) + pI_x(p+1,q+1) = 0.$$

Notice that the sum of the coefficients is always zero.

By a repeated use of (10) it is possible to obtain the formulae

$$(15)\quad I_x(p+n,q) = \frac{1}{(p+n-1)^{(n)}}\sum_{r=0}^{n}(-1)^r \cdot \binom{n}{r}(p+q+n-1)^{(n-r)}(q+r-1)^{(r)}I_x(p,q+r),$$

$$(16)\quad I_x(p,q+n) = \frac{1}{(q+n-1)^{(n)}}\sum_{r=0}^{n}(-1)^r \cdot \binom{n}{r}(p+q+n-1)^{(n-r)}(p+r-1)^{(r)}I_x(p+r,q),$$

where $(p+q+n-1)^{(n-r)}$, etc., refer to the factorial notation, e. g.,

$$[p+q+(n-1)]^{(n-r)} = (p+q+n-1)(p+q+n-2)\cdots(p+q+r).$$

3. An application. Formulae (15) and (16) may be used to write general formulae for obtaining values of $I_x(p,q)$ where p or q may be greater than 50, i. e., for such values outside the range of Karl Pearson's tables. In particular,

$$(17)\quad I_x(50+n,q) = \frac{1}{(49+n)^{(n)}}\left[(n+q+49)^{(n)}I_x(50,q) - \binom{n}{1}q(n+q+49)^{(n-1)}I_x(50,q+1)\cdots(-1)^n(q+n-1)^{(n)}I_x(50,q+n)\right]$$

and

$$(18)\quad I_x(p,50+n) = \frac{1}{(49+n)^{(n)}}\left[(n+p+49)^{(n)}I_x(p,50) - \binom{n}{1}p(n+p+49)^{(n-1)}I_x(p+1,50)\cdots(-1)^n(p+n-1)^{(n)}I_x(p+n,50)\right].$$

It should be noted for (17) that as n increases the range of values that can be obtained outside Karl Pearson's tables are reduced since the last term of (17) contains $I_x(50, q+n)$. A similar observation is noted for (18). From a practical standpoint the computational labor restricts n to fairly small values. Using (17) we may easily compute for example,

$$I_{.60}(52,48) = I_{.60}(50+2,48)$$

$$= \frac{1}{(51)(50)}[(99)(98)I_{.60}(50,48) - 2(99)(48)I_{.60}(50,49) + (49)(48)I_{.60}(50,50)].$$

Substituting the necessary values from Karl Pearson's tables we calculate

$$I_{.60}(52, 48) = .9465248.$$

Similarly using (18) we may calculate

$$I_{.40}(48, 52) = .0534752.$$

As a check on the computations, we use the well-known identity

$$I_x(p, q) = 1 - I_{1-x}(p', q'),$$

where $p' = q$ and $q' = p$. Then

$$\begin{aligned} I_{.40}(48, 52) &= 1 - I_{.60}(52, 48) \\ &= 1 - .9465248 \\ &= .0534752. \end{aligned}$$

In like manner formulae (15) and (16) may be used to write general formulae for obtaining half values for p or q greater than 10.5, i. e., for values not included in Karl Pearson's tables. In particular,

$$\begin{aligned} I_x(10.5 + n, q) = \frac{1}{(9.5 + n)^{(n)}} \Big[&(9.5 + q + n)^{(n)} I_x(10.5, q) - \binom{n}{1} \\ &\cdot q(9.5 + q + n)^{(n-1)} I_x(10.5, q + 1) \cdots (-1)^n (q + n - 1)^{(n)} I_x(10.5, q + n) \Big], \end{aligned} \tag{19}$$

and

$$\begin{aligned} I_x(p, 10.5 + n) = \frac{1}{(9.5 + n)^{(n)}} \Big[&(9.5 + p + n)^{(n)} I_x(p, 10.5) - \binom{n}{1} \\ &\cdot p(9.5 + p + n)^{(n-1)} I_x(p + 1, 10.5) \cdots (-1)^n (p + n - 1)^{(n)} I_x(p + n, 10.5) \Big]. \end{aligned} \tag{20}$$

Using (19) we may compute

$$\begin{aligned} I_{.60}(12.5, 8) = \frac{1}{(11.5)^{(2)}} [&(19.5)^{(2)} I_{.60}(10.5, 8) - 2(8)(19.5) I_{.60}(10.5, 9) \\ &+ (9)(8) I_{.60}(10.5, 10)], = .4512367. \end{aligned}$$

Similarly using (20) we obtain

$$I_{.40}(8, 12.5) = .5487633.$$

Employing the check formula,

$$\begin{aligned} I_{.40}(8, 12.5) &= 1 - I_{.60}(12.5, 8) \\ &= 1 - .4512367 \\ &= .5487633. \end{aligned}$$

Thanks are due to Dr. J. C. P. Miller, Technical Director, Scientific Computing Service, Limited, London, England, for helpful suggestions in the preparation of this paper.

REFERENCES

[1] T. A. Bancroft, "Some extensions of the incomplete beta function tables." (in preparation)

[2] Karl Pearson, *Tables of the Incomplete Beta-Function*, Cambridge University Press, 1934.

[3] Wilhelm Magnus und Fritz Oberhettinger, *Formeln and Sätze für die Speziellen Funktionen der Mathematischen Physik*, Julius Springer, Berlin, 1943.

[4] T. A. Bancroft, "On biases in estimation due to the use of preliminary tests of significance, *Annals of Math. Stat.*," Vol. 15 (1944).

ON A THEOREM BY WALD AND WOLFOWITZ

By Gottfried E. Noether

New York University

Let $\mathfrak{H}_n = (h_1, \cdots, h_n)$, $(n = 1, 2, \cdots)$, be sequences of real numbers and for all n denote by $H_{e_1 \cdots e_m}$ the symmetrical function generated by $h_1^{e_1} \cdots h_m^{e_m}$, i.e., $H_{e_1 \cdots e_m} = \Sigma h_{i_1}^{e_1} \cdots h_{i_m}^{e_m}$ where the summation is extended over the $n(n-1) \cdots (n-m+1)$ possible arrangements of the m integers $i_1, \cdots, i_m$, such that $1 \leq i_j \leq n$ and $i_j \neq i_k$, $(j, k = 1, \cdots, m)$. According to Wald and Wolfowitz [1] the sequences $\mathfrak{H}_n$ are said to satisfy condition W, if for all integral $r > 2$

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(h_i - \bar{h})^r}{\left[\frac{1}{n}\sum_{i=1}^{n}(h_i - \bar{h})^2\right]^{r/2}} = O(1),^1$$

where $\bar{h} = 1/n \, \Sigma_{i=1}^{n} h_i$.

Given sequences $\mathfrak{A}_n = (a_1, \cdots, a_n)$ and $\mathfrak{D}_n = (d_1, \cdots, d_n)$, consider the chance variable

$$L_n = d_1 x_1 + \cdots + d_n x_n,$$

where the domain of $(x_1, \cdots, x_n)$ consists of the $n!$ equally likely permutations of the elements of $\mathfrak{A}_n$. Then it is shown in [1] that if the sequences $\mathfrak{A}_n$ and $\mathfrak{D}_n$ satisfy condition W, the distribution of $L_n^0 = (L_n - EL_n)/\sigma(L_n)$ approaches the normal distribution with mean 0 and variance 1 as $n \to \infty$. These conditions

[1] The symbol O, as well as the symbols o and $\sim$ to be used later, have their usual meaning. See e. g. Cramér [2, p. 122].

for asymptotic normality can be weakened. It will be shown that the following theorem holds:

THEOREM. *L_n^0 is asymptotically normal with mean 0 and variance 1 provided the sequences $\mathfrak{D}_n$ satisfy condition W while for the sequences $\mathfrak{A}_n$*

$$(1) \qquad \frac{\sum_{i=1}^{n} (a_i - \bar{a})^r}{\left[\sum_{i=1}^{n} (a_i - \bar{a})^2\right]^{r/2}} = o(1), \qquad (r = 3, 4, \cdots).$$

We note that L_n^0 is not changed if a_i is replaced by $[1/n\Sigma_{i=1}^{n} (a_i - \bar{a})^2]^{-1/2}(a_i - \bar{a})$ and d_i by $[1/n \Sigma_{i=1}^{n} (d_i - \bar{d})^2]^{-1/2}(d_i - \bar{d})$. Therefore it is sufficient to prove asymptotic normality provided

$$(2) \qquad D_1 = 0, \quad D_2 = n, \quad D_r = O(n), \qquad (r = 3, 4, \cdots);$$

$$(3) \qquad A_1 = 0, \quad A_2 = n, \quad A_r = o(n^{r/2}), \qquad (r = 3, 4, \cdots).$$

Then

$$\begin{aligned} EL_n &= D_1 Ex_1 = 0, \\ \operatorname{var} L_n = EL_n^2 &= D_2 Ex_1^2 + D_{11} Ex_1 x_2 \\ &= \frac{1}{n} A_2 D_2 + \frac{1}{n(n-1)} (A_1^2 - A_2)(D_1^2 - D_2) \sim n, \end{aligned}$$

and it is sufficient to show that $n^{-r/2} EL_n^r$ tends to the rth moment of a normal distribution with mean 0 and variance 1.

Now we can write

$$\begin{aligned} \mu_r = n^{-r/2} EL_n^r &= n^{r/2} \sum_{i_1=1}^{n} \cdots \sum_{i_r=1}^{n} E d_{i_1} x_{i_1} \cdots d_{i_r} x_{i_r} \\ (4) \qquad &= n^{-r/2} [D_r Ex_1^r + \cdots + c(r, e_1, \cdots, e_m) D_{e_1 \cdots e_m} Ex_1^{e_1} \cdots x_m^{e_m} \\ &\qquad + \cdots + D_{1 \cdots 1} Ex_1 \cdots x_r] \end{aligned}$$

where $e_1 + \cdots + e_m = r$ with e_k, $(k = 1, \cdots, m)$, positive integral and the coefficient $c(r, e_1, \cdots, e_m)$ stands for the number of ways in which the r indices $i_1, \cdots, i_r$ can be tied in m groups of size $e_1, \cdots, e_m$, respectively, so as to produce the terms of $D_{e_1 \cdots e_m} Ex_1^{e_1} \cdots x_m^{e_m}$.

Since $Ex_1^{e_1} \cdots x_m^{e_m} \sim n^{-m} A_{e_1 \cdots e_m}$ we have

$$(5) \quad n^{-r/2} D_{e_1 \cdots e_m} Ex_1^{e_1} \cdots x_m^{e_m} \sim n^{-(r/2+m)} D_{e_1 \cdots e_m} A_{e_1 \cdots e_m} = B(r, e_1, \cdots, e_m), \text{ say.}$$

LEMMA. *$B(r, e_1, \cdots, e_m) \sim 0$ unless*

$$(6) \qquad m = r/2, \qquad e_1 = \cdots = e_{r/2} = 2.$$

In that case $B(r, 2, \cdots, 2) \sim 1$.

Before proving this lemma we shall show that our theorem follows immediately. By (4) μ_r is the sum of a finite number of expressions $B(r, e_1, \cdots, e_m)$.

Therefore if $r = 2s + 1$, $(s = 1, 2, \cdots)$, $\mu_{2s+1} \sim 0$, since at least one of the e_k, $(k = 1, \cdots, m)$, in all the $B(2s + 1, e_1, \cdots, e_m)$ adding up to μ_{2s+1} must be odd. If $r = 2s$, $\mu_{2s} \sim c(2s, 2, \cdots, 2)$. Since the first index in (4) can be tied with any one of the other $2s - 1$ indices, the next free index with any one of the remaining $2s - 3$ indices, etc., it is seen that $\mu_{2s} \sim (2s - 1)(2s - 3) \cdots 3$. However these are the moments of a normal distribution with mean 0 and variance 1. This proves the theorem.

PROOF OF LEMMA. Define $A(j_1, \cdots, j_h) = A_{j_1} \cdots A_{j_h}$. Then $A_{e_1 \cdots e_m}$ is the sum of a finite number of expressions $A(j_1, \cdots, j_h)$, where the j_g, $(g = 1, \cdots, h)$, are obtained from $e_1, \cdots, e_m$ by addition in such a way that

$$j_1 + \cdots + j_h = e_1 + \cdots + e_m = r. \tag{7}$$

Since by (3) $A_1 = 0$, we need only consider those $A(j_1, \cdots, j_h)$ for which $j_g \geq 2$, $(g = 1, \cdots, h)$. If some $j_g > 2$ by (3) and (7)

$$A(j_1, \cdots, j_h) = o(n^{r/2}). \tag{8}$$

If $j_g \equiv 2$,

$$A(2, \cdots, 2) = A_2^{r/2} = n^{r/2}. \tag{9}$$

This last case can only happen if r is even and e_k, $(k = 1, \cdots, m)$, equals either 1 or 2. Therefore, unless (6) is true

$$m > r/2. \tag{10}$$

Similarly, writing $D_{e_1 \cdots e_m}$ as a sum of products of the kind $D_{j_1} \cdots D_{j_h}$ it is seen that by (2)

$$D_{e_1 \cdots e_m} = \begin{cases} O(n^m) & \text{if } m < r/2 \\ O(n^{r/2}) & \text{if } m \geq r/2. \end{cases} \tag{11}$$

Thus by (8)–(11)

$$A_{e_1 \cdots e_m} D_{e_1 \cdots e_m} = o(n^{r/2+m}), \tag{12}$$

unless (6) is true. In that case

$$A_{2 \cdots 2} \sim A_2^{r/2} = n^{r/2}, \tag{13}$$

$$D_{2 \cdots 2} \sim D_2^{r/2} = n^{r/2}. \tag{14}$$

(12)–(14) together with (5) prove the lemma.

Let $a_1, a_2, \cdots$ be independent observations on the same chance variable Y. We may ask what conditions have to be imposed on the distribution of Y to insure—at least with probability 1—that condition (1) is satisfied. Wald and Wolfowitz state in Corollary 2 of [1] that provided Y has positive variance and finite moments of all orders the $a_1, a_2, \cdots$ satisfy condition W with probability 1 and therefore insure asymptotic normality of L_n provided the sequences $\mathfrak{D}_n$ satisfy condition W. On the other hand, it can be shown that the $a_1, a_2, \cdots$

satisfy condition (1) with probability 1, provided Y has positive variance and a finite absolute moment of order 3. Thus condition (1) constitutes a considerable improvement over condition W.

REFERENCES

[1] A. WALD AND J. WOLFOWITZ, "Statistical tests based on permutations of the observations," *Annals of Math. Stat.*, Vol. 15 (1944), pp. 358-372.
[2] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton, 1946.

ON SUMS OF SYMMETRICALLY TRUNCATED NORMAL RANDOM VARIABLES

BY Z. W. BIRNBAUM AND F. C. ANDREWS[1]
University of Washington, Seattle

1. Introduction. Let X_a be the random variable with the probability density

$$f_a(x) = \begin{cases} Ce^{-x^2/2} & \text{for } |x| \leq a \\ 0 & \text{for } |x| > a, \end{cases} \tag{1.1}$$

obtained from the normal probability density $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ by symmetrical truncation at the "terminus" $|x| = a$, and let $S_a^{(m)}$ be the sum of m independent sample-values of X_a. We consider the following *problem*: An integer $m \geq 2$ and the real numbers $A > 0$, $\epsilon > 0$ are given; how does one have to choose the terminus a so that the probability of $|S_a^{(m)}| \geq A$ is equal to ϵ,

$$P(|S_a^{(m)}| \geq A) = \epsilon? \tag{1.2}$$

This problem arises for example when single components of a product are manufactured under statistical quality control, so that each component has the length $Z = k + X$ where X has the probability density $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, and the final product consists of m components so that its total length S is the sum of the lengths of the components. We wish to have probability $1 - \epsilon$ that S differs from mk by not more than a given A. To achieve this we decide to reject each single component for which $|Z - k| = |X| > a$; how do we determine a?

The exact solution of this problem would require laborious computations.[2] In the present paper methods are given for obtaining approximate values of a which are "safe", that is such that

$$P(|S_a^{(m)}| \geq A) \leq \epsilon. \tag{1.3}$$

[1] Research done under the sponsorship of the Office of Naval Research.

[2] A similar problem has been studied by V. J. Francis [2] for one-sided truncation; he actually had the exact probabilities for the solution of his problem computed and tabulated for $m = 2, 4$.

In deriving these safe values, use will be made of theorems on random variables with comparable peakedness, for which the reader is referred to a previous paper [1].

2. The safe value a_1. For fixed $a > 0$, we consider the normal random variable Y_a with expectation 0 and with probability density $g_a(Y_a)$ such that $g_a(0) = f_a(0)$. It is easily seen that Y_a has the standard deviation

$$\sigma_a = \frac{1}{\sqrt{2\pi}} \int_{-a}^{+a} e^{-t^2/2}\, dt, \tag{2.1}$$

and that $g_a(\xi) < f_a(\xi)$ for $|\xi| \leq a$, $g_a(\xi) > 0 = f_a(\xi)$ for $|\xi| > a$. Hence, applying Theorem 1 in [1], we conclude that

$$P(|S_a^{(m)}| \geq A) \leq \frac{2}{\sqrt{2\pi}} \int_{(A/\sigma_a\sqrt{m})}^{\infty} e^{-t^2/2}\, dt. \tag{2.2}$$

If m, A, and ϵ are given, we determine ξ_ϵ from tables of the normal probability integral so that $\frac{2}{\sqrt{2\pi}} \int_{\xi_\epsilon}^{\infty} e^{-t^2/2}\, dt = \epsilon$, set $\sigma_a = \frac{A}{\xi_\epsilon\sqrt{m}}$ in (2.1), and solve the equation

$$\frac{A}{\xi_\epsilon\sqrt{m}} = \frac{1}{\sqrt{2\pi}} \int_{-a}^{+a} e^{-t^2/2}\, dt \tag{2.3}$$

for a using again tables of the normal probability integral. In view of (2.2) this solution satisfies (1.3) and hence is safe; it will be denoted by a_1.

3. The safe value a_2. A direct application of Theorem 2 in [1] yields the inequality

$$P(|S_a^{(m)}| \geq A) \leq \frac{1}{2^{m-1}m!} \sum_{\frac{1}{2}(m+A/a) < j \leq m}^{m} (-1)^j \binom{m}{j} \left(\frac{A}{a} + m - 2j\right)^m = h_m\left(\frac{A}{a}\right) \tag{3.1}$$

for $0 \leq A \leq ma$. Hence by equating $h_m(A/a)$ to ϵ and solving for a, we obtain a safe value which will be denoted by a_2. It is of interest to note that (3.1) *is true* not only for $f_a(x)$ defined by (1.1) i.e. truncated normal, but *for any probability density $f_a(x)$ which is symmetrical and unimodal*, since these are the only assumptions needed for Theorem 2 in [1].

4. Solution for large m. The random variable X_a has the variance

$$\sigma^2(X_a) = 1 + \frac{2\phi''(a)}{2\phi(a) - 1} \tag{4.1}$$

where

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

Hence, according to the central limit theorem, we have the approximate equality

$$P(|S_a^{(m)}| \geq A) = \frac{2}{\sqrt{2\pi}} \int_{(A/\sigma(X_a)\sqrt{m})}^{\infty} e^{-t^2/2}\, dt \tag{4.2}$$

for m sufficiently large.

It can be reasonably expected that the cumulative distribution of $S_a^{(m)}$ differs from its limiting normal probability integral by less than the cumulative distribution of the sum $U_a^{(m)}$ of m independent uniform variables in $(-a, +a)$ differs from its limiting normal probability integral. Already for $m = 4$ the cumulative distribution of $U_a^{(m)}$ differs from the corresponding normal cumulative by less than .0075. Equally good or better approximation may, therefore, be expected for the distribution of $S_a^{(m)}$, so that the error in the approximate equality (4.2) between the two-tail probabilities should be less than .015 for $m = 4$, and still less for $m > 4$.

Equating the right-hand term of (4.2) to ϵ and solving for $\sigma^2(X_a)$, we obtain

$$\sigma^2(X_a) = 1 + \frac{2\phi''(a)}{2\phi(a) - 1} = \frac{1}{m}\left(\frac{A}{\xi_\epsilon}\right)^2,$$

an equation which can be solved for a with the aid of tables of $\phi(x)$ and $\phi''(x)$. We denote this value of a by α_1.

5. Use of the different solutions in practice. From the foregoing it appears that the following procedure may be followed in solving our problem in any definite case:

If m is large, α_1 is very close to the exact solution of (1.3) and may be used safely.

If m is not large but $m \geq 5$, it is conjectured that α_1 is such that the left-hand term in (1.3), for $a = \alpha_1$, differs from ϵ by less than 0.015.

If $m \leq 4$, the larger of a_1 and a_2 should be used. Table I contains the A for which a_1 and a_2 have the same value, say a'; a_1 or a_2 should be used if the given A is greater or smaller, respectively, than the tabulated value. The value a_1 is easily computed from a table of the normal probability integral by the procedure of section 2. The value a_2 can be obtained by reading off A/a_2 from Table II.

TABLE I

Values of A for which $a_1 = a_2 = a'$ for given m, ϵ

ϵ \ m	2		3		4	
	A	a'	A	a'	A	a'
.001	4.568	2.357	5.446	2.008	6.152	1.842
.002	4.258	2.228	5.059	1.918	5.717	1.779
.005	3.808	2.047	4.512	1.799	5.111	1.697
.01	3.438	1.910	4.074	1.712	4.632	1.640
.02	3.034	1.765	3.614	1.630	4.131	1.589
.05	2.456	1.581	2.970	1.533	3.425	1.529

TABLE II

Values of A/a_2 for given m, ϵ

ϵ \ m	2	3	4
.001	1.937	2.712	3.339
.002	1.911	2.637	3.213
.005	1.859	2.507	3.011
.01	1.800	2.379	2.824
.02	1.718	2.217	2.600
.05	1.553	1.937	2.240

6. Examples. 1) $A = 3.8$, $m = 4$, $\epsilon = .05$. Since A is greater than the value 3.425 in Table I, we compute $a_1 = 2.162$. From Table II we would obtain $A/a_2 = 2.240$ and thus $a_2 = 1.696 < a_1$. 2) $A = 3$, $m = 4$, $\epsilon = .02$. Since $A < 4.131$, we read $A/a_2 = 2.600$ from Table II and obtain $a_2 = 1.153$ which will be greater than a_1. 3) $A = 5$, $m = 30$, $\epsilon = .05$. Using the method of section 4 we obtain $\alpha_1 = 1.62$.

REFERENCES

[1] Z. W. Birnbaum, "On random variables with comparable peakedness," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 76–81.

[2] V. J. Francis, "On the distribution of the sum of n sample values drawn from a truncated normal population," *Roy. Stat. Soc. Jour. Suppl.*, Vol. 8 (1946), pp. 223–232

A CERTAIN CUMULATIVE PROBABILITY FUNCTION

By Sister Mary Agnes Hatke, O.S.F.

St. Francis College, Ft. Wayne, Indiana

Graduations of empirically observed distributions show that the cumulative probability function $F(x) = 1 - (1 + x^{1/c})^{-1/k}$ is a practical tool for fitting a smooth curve to observed data. The graduations are comparable with those obtained by the Pearson system, Charlier, and others and are accomplished with simple calculations. Given distributions are graduated by the method of moments. Theoretical frequencies are obtained by evaluation of consecutive values of $F(x)$ by use of calculating machines and logarithms, and by differencing $NF(x)$. No integration nor heavy interpolation is involved, such as may be required in graduation by a classical frequency function. Burr [1] constructed tables of ν_1, σ, α_3, and α_4 values for the function $F(x)$ for certain combinations of integral values of $1/c$ and $1/k$. In these tables curvilinear interpolation must be used in finding an $F(x)$ with desired moments. The writer constructed more extensive tables for the same cumulative function with c and k a variety of real positive numbers less than or equal to one, such that linear interpolation can be used to determine the parameters c and k for an $F(x)$ that has α_3 and α_4 approximately the same as those of the distribution to be graduated. These tables have been deposited with Brown University. Microfilm or photostat copies may be obtained upon request to the Brown University Library.

The writer used the definitions of cumulative moments and the formulas for the ordinary moments ν_1, σ, α_3, and α_4 in terms of cumulative moments as developed by Burr. These latter moments were tabulated for the function $F(x)$ having various combinations of parameters c and k, c ranging from 0.050 to 0.675 and k from 0.050 to 1.000, each at intervals of 0.025. Within these ranges only those combinations of c and k were used which yielded α_3 of approximately 1 or less and α_4 values of 6 or less, since such moments are most common in practice.

It can be verified that over most of the area of the table α_3 values obtained

by linear and by curvilinear interpolation on k (or on c) differ by less than 0.001 and values of α_4 by approximately 0.01 or less. If α_3 = constant and α_4 = constant curves are plotted on c, k axes, it will be seen that there exists only one solution (c, k) of the equations $\alpha_3 = B(c, k)$ and $\alpha_4 = C(c, k)$. Furthermore, some α_4 curves intersect two α_3 curves representing the same $|\alpha_3|$. Thus the chance of finding an appropriate function $F(x)$ for graduation is increased since by reversal of scale an $F(x)$ with a positive α_3 may be used to graduate a distribution with a negative α_3, and conversely.

Graduation of an observed frequency distribution is easily accomplished. Linear interpolation on k for a fixed c seems to be the best method for determining

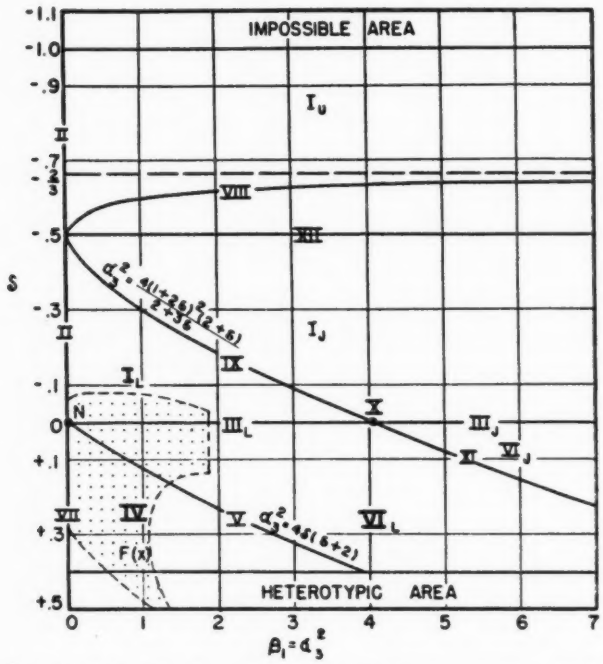

FIG. 1. The α_3^2, δ chart for the Pearson system of frequency curves and the area covered by $f(x) = 1 - (1 + x^{1/c})^{-1/k}$ (subscript L = bell-shaped)

the parameters of an $F(x)$ that has α_3 exactly the same and α_4 nearly the same as the observed α_3 and α_4. If the observed α_3 and α_4 are fairly close to an entry in the table, no interpolation is required. Direct linear interpolation is used to determine ν_1 and σ for the c and k just found. Letting M and S be the mean and standard deviation of the given distribution, the formula,

$$\frac{x - \nu_1}{\sigma} = t = \frac{X - M}{S}$$

is used to translate the class limits X of the given distribution to the corresponding x's of $F(x)$. For any x that is negative the quantity $1 + x^{1/c}$ is taken as one

to make $F(-x) = 0$ in accordance with the definition of $F(x)$ [1]. The values of $(1 + x^{1/c})^{-1/k}$ for the various x's are computed by logarithms and differenced to obtain the probabilities for the given class intervals, according to equation

$$P(a \leq x \leq b) = \int_a^b f(x)\,dx = F(b) - F(a).$$

The respective theoretical frequencies are these probabilities multiplied by N, the number of cases.

The headings that proved satisfactory for the columns of the graduation work-sheet are: class intervals (in observed physical units), X (u if unit class-interval is used), f_{obs}, x, $1 + x^{1/c}$, $N/(1 + x^{1/c})^{1/k}$, and f_{tk}.

The relation of $F(x)$ to the Pearson system of frequency curves is presented in Figure 1, which is a reproduction of a major part of Craig's chart for α_3^2 and δ [2]. In this chart the parameters of the twelve Pearson curves are expressed in terms of α_3^2 and δ, where $\delta = (2\,\alpha_4 - 3\alpha_3^2 - 6)/(\alpha_4 + 3)$. Values of α_3^2 and δ were computed for $F(x) = 1 - (1 + x^{1/c})^{-1/k}$ in which c and k were assigned the values listed in the α_3, α_4 table. The dotted area superimposed on the Craig chart is that covered by these α_3^2, δ values for $F(x)$. Although it is small in size compared to the total area, it contains a part of the areas representing the three main Pearson curves, I, IV, and VI, as well as the point for the normal curve and part of the line on which lie the points corresponding to the bell-shaped curves of the Type III functions. It also includes transitional Types V and VII. Thus the function $F(x)$ covers part of an important area on the α_3^2, δ chart for the Pearson curves.

The function $F(x)$ was used to graduate satisfactorily several observed distributions classified as Pearson types, including the three main Types, I, IV, and VI, and transitional Types III and VII.

One advantage in the use of this cumulative function $F(x)$ is that it takes but one symbolic form with the area covered, whereas the Pearson-system curves require several different expressions of various complexity requiring identification of type. Furthermore, graduation by a Pearson function generally involves approximate integration or heavy interpolation in the incomplete beta function tables for the evaluation of the integrals of the Pearson functions, whereas graduation by a function $F(x)$ is easily and quickly performed since $F(x)$ only involves two number-parameters readily determined by means of the α_3, α_4 table and straight arithmetic.

The writer is deeply indebted to Professor Irving W. Burr of Purdue University for valuable suggestions in this study.

REFERENCES

[1] I. W. Burr, "Cumulative frequency functions," *Annals of Math. Stat.*, Vol. 13 (1942), pp. 215–232.

[2] C. C. Craig, "A new exposition and chart for the Pearson system of frequency curves," *Annals of Math. Stat.*, Vol. 7 (1936), pp. 16–28.

ABSTRACTS OF PAPERS

(Presented at the Berkeley Meeting of the Institute, June 16–18, 1949)

1. **Extension of a Theorem of Blackwell.** E. W. Barankin, University of California, Berkeley.

It is proved that Blackwell's method of uniformly improving the variance of an unbiased estimate by taking the conditional expectation with respect to a sufficient statistic, is, in fact, similarly effective on every absolute central moment of order $s \geqq 1$. The method leads to finer detail concerning the relationship between an estimate and its thus derived one. (This paper was prepared with the partial support of the Office of Naval Research.)

2. **On the Existence of Consistent Tests.** Agnes Berger, Columbia University., New York.

Let $\mathfrak{M}(\mathfrak{B})$ denote the space of all probability-measures defined over a common Borel-field $\mathfrak{B}$. Let $\{m\} = M$, $\{m'\} = M'$ be two disjoint subsets of $\mathfrak{M}(\mathfrak{B})$ and let H_0 (H_1) be the hypothesis stating that the unknown distribution is in M (M'). In Neyman's terminology H_0 can be consistently tested against H_1 if to any preassigned $\epsilon > 0$ there exists an integer n and a critical region in the product-space of n independent observations such that the probabilities of the errors of the first and second kind corresponding to this region are simultaneously smaller than ϵ. A sufficient condition which for a certain type of consistent test is also necessary is established. The condition is satisfied whenever the disjoint sets M and M' are closed and compact with respect to a certain suitable topology introduced on $\mathfrak{M}(\mathfrak{B})$. Thus for instance H_0 can be consistently tested against H_1 if M and M' contain only a finite number of measures or if the measures in M resp. M' depend continuously on a parameter ranging over a closed and bounded subset of some Euclidean space.

3. **Effect of Linear Truncation in a Multinormal Population.** Z. William Birnbaum, University of Washington, Seattle.

Let $(X, Y_1, Y_2, \cdots, Y_{n-1})$ have a non-singular n-dimensional normal probability density $f(X, Y_1, Y_2, \cdots, Y_{n-1})$ for which all parameters are given, and let $\varphi(X, Y_1, Y_2, \cdots, Y_{n-1})$ be the probability density obtained from f by truncation along a given hyperplane: $\varphi = Cf$ for $a_1 Y_1 + \cdots + a_{n-1} Y_{n-1} \leqq aX + b$, $\varphi = 0$ elsewhere. What is the marginal distribution of X for this truncated distribution? This question can be answered by using a set of tables with only two parameters. These tables make it also possible to solve problems such as: determine the plane of truncation so that the marginal distribution of X has certain required proper ·es. (This paper was prepared under the sponsorship of the Office of Naval Research.)

4. **Statistical Problems in the Theory of Counters.** (Preliminary Report). Colin R. Blyth, University of California, Berkeley.

The assumptions made about counter action and distribution of incident particles are the same as those of B. V. Gnedenko [On the theory of Geiger-Müller counters, *Journ. Exper. i Teor. Phiz*, Vol. 11 (1941)]. The distribution of the number X of particles registered during a given time $(0, t)$ is found explicitly, in terms of the density $a(v)$ of incident particles at time v. The problem considered is that of estimating the parameters of $a(v)$. For the special case $a(v) = a$, the distribution of X reduces to $P\{X = x\} = a^x(t - x\tau)^x \exp$

$\{-a(t-x\tau)\}/x! + \exp\{-a(t-x\tau)\}\Sigma_{i=0}^{x-1} a^i[t-x\tau]^i/i! - \exp\{-a[t-(x-1)\tau]\}\Sigma_{i=0}^{x-1} a^i[t-(x-1)\tau]^i/i!$ for $x = 1, 2, \cdots, s = \left[\frac{t}{\tau}\right]$; $P\{X=0\} = e^{-at}$; $P\{X = s+1\} = 1 - \exp\{-a(t-s\tau)\}\sum_{i=0}^{s} a^i[t-s\tau]^i/i!$; $P\{X > s+1\} = 0$. This distribution has been found in another problem by J. Neyman [*On the problem of estimating the number of schools of fish*, submitted to Statistical Series, Univ. of Calif. press]. For this special case the maximum likelihood estimate $\hat{a}$ of a is found to be given by $\hat{a}\tau \exp(\hat{a}\tau) = \{1 + \tau/(t - x\tau)\}^x x\tau/(t - x\tau)$. If $\tau/(t - x\tau)$ is small, as will usually be the case, $\hat{a}$ will be close to the estimate $x/(t - x\tau)$ usually used for a.

5. **Some Two-Sample Tests.** Douglas G. Chapman, University of California, Berkeley.

Let X, Y be random variables normally distributed with means ξ, η, variances σ_1, σ_2 respectively. The two sample procedure formulated by Stein to obtain a test with power independent of σ, for the hypothesis $\eta = \xi_0$ is used here to determine a test for the hypothesis $\frac{\xi}{\eta} = r$ (r any pre-assigned real number). The size and power of this test are independent of σ_1 and σ_2. The two sample procedure may be extended to the more general case of testing the hypothesis of equality of means of several normal populations, the variances being unknown. Approximate tests are obtained for this case. Finally it is shown that this two sample procedure can be used to select that normal population, of several, with the greatest mean: the rule of selection having a preassigned level of accuracy. (This paper was prepared with the partial support of the Office of Naval Research.)

6. **Minimum Variance in Non-Regular Estimation.** R. C. Davis, U. S. Naval Ordnance Test Station, Inyokern.

The Cramér-Rao inequality for the minimum variance of a regular estimate of an unknown parameter of a probability distribution is extended to a broad class of non-regular types of estimation. The theory is developed only for the case in which a probability density function and a sufficient statistic for the unknown parameter exist. For every non-regular estimation problem included in the above class, it is proved that there exists a unique unbiased estimate which attains minimum variance, and a method is given for obtaining the sample estimate. Examples are given; such as, the rectangular distribution, a class of truncated distributions, etc.

7. **Auxiliary Random Variables.** Mark W. Eudey, California Municipal Statistics, Inc., San Francisco.

In testing hypotheses concerning discontinuous random variables it is not possible to find regions of arbitrary size, and so if we compare two critical regions, selection between them on the basis of the usual criteria of the Neyman-Pearson theory of testing hypotheses may be confused by the difference in their sizes. This difficulty may be avoided by allowing the statistician to use a mixed strategy in such cases, and make his decision to accept or reject the hypothesis depend upon an independent auxiliary random variable. For example, if K is a binomial variable, and U has a uniform distribution $(0, 1)$, then $Z = K + U$ may be used to test hypotheses concerning the binomial parameter, and regions of any size may be found. For the binomial case this procedure leads to a class of uniformly most powerful tests for one-sided alternatives, and to uniformly most powerful unbiased tests for two-

sided alternatives. Similar results are obtained for other common discontinuous variables, and the same device may be used in considering confidence regions and decision functions for such variables. (This paper was prepared with the partial support of the Office of Naval Research.)

8. **Estimation in Truncated Samples.** MAX HALPERIN, The Rand Corporation. Santa Monica, California.

A death process is considered which starts with n individuals of zero age, each following the mortality law, $f(x, \theta)$. That is,

$$F(t) = Pr\ \{\text{Age at death} < t\} = \int_0^t f(x, \theta)\, dx,$$

where $f(x, \theta)$ is a probability density. We suppose we truncate the process at a fixed time, T, and wish to estimate θ when

a) individuals who die are not replaced, and
b) individuals who die are replaced by individuals of zero age following the mortality law, $f(x, \theta)$.

In both cases, it is found that, under mild conditions, estimation by Maximum Likelihood gives optimum estimates. The estimates are best in the sense of being asymptotically normally distributed and of minimum variance for large samples.

The proofs are given for the case of a single parameter, but can be extended to the multiparameter case. Examples are given.

9. **Some Problems in Point Estimation.** J. L. HODGES, JR. AND E. L. LEHMANN, University of California, Berkeley.

Some point estimation problems are considered in the light of Wald's general theory. It is shown that when the loss function is convex, one may restrict consideration to nonrandomized estimates based on sufficient statistics. Minimax estimates are obtained in a number of cases connected with the binomial and hypergeometric distributions, and with some non-parametric problems. Some prediction problems are also considered. (This paper was prepared with the partial support of the Office of Naval Research.)

10. **Completeness in the Sequential Case.** E. L. LEHMANN AND C. STEIN, University of California, Berkeley.

Recently, in a series of papers, Girshick, Mosteller, Savage and Wolfowitz have considered the uniqueness of unbiased estimates depending only on an appropriate sufficient statistic for sequential sampling schemes of binomial variables. A complete solution was obtained under the restriction to bounded estimates. This work, which has immediate consequences with respect to the existence of unbiased estimates with uniformly minimum variance, is extended here in two directions. A general necessary condition for uniqueness is found, and this is applied to obtain a complete solution of the uniqueness problem when the random variables have a Poisson or rectangular distribution. Necessary and sufficient conditions are also found in the binomial case without the restriction to bounded estimates. This permits the statement of a somewhat stronger optimum property for the estimates, and is applicable to the estimation of unbounded functions of the unknown probability.

11. **The Ratio of Ranges.** RICHARD F. LINK, University of Oregon, Eugene.

The distribution of the ratio of two ranges from independent samples drawn from a normal population is given analytically for n_1 and $n_2 \leqq 3$. A table of percentage values, R,

is given for $\alpha = .005, .01, .025, .05, .10$ and for all combinations of n_1 and n_2 up to 10, where $\alpha = \Pr(w_1/w_2 > R)$ and w_1 and w_2 are the observed ranges. (This paper was prepared under the sponsorship of the Office of Naval Research.)

12. **Some Problems Arising in Plant Selection and the Use of Analysis of Variance.** STANLEY W. NASH, University of California, Berkeley.

The yields of many (m) varieties are compared in a field trial. A few varieties having the highest and lowest yields in this trial are selected for further testing. What chance is there that the first trial will give a significant result, the second trial not? Let ξ_i denote the true mean yield of the ith variety, and assume that the ξ_i are themselves normally, independently distributed with variance σ_1^2. Let P_k $(k = 1, 2)$ denote the probability of a significant result in the kth trial, using the F-test. For fixed $\sigma_1^2 > 0$, $\lim_{m\to\infty} P_1 = 1$. (See Nash, *Annals of Math. Stat.*, Vol. 19 (1948), p. 434.) Now let $\sigma_1^2 > 0$ take on a decreasing sequence of values as m increases. If $\frac{1}{\sigma_1^2 g(m)} = 0\left(\frac{E(F)}{\sigma_F}\right)$, then $\lim_{m\to\infty} P_1 = 1$. Here $1 + \sigma_1^2 g(m) = \frac{E(\text{numerator of } F)}{\sigma_0^2\ (= \text{error variance})}$. Also $\lim_{m\to\infty} P_2 < 1$ if and only if $\sigma_1^2 = 0\left(\frac{1}{\sqrt{\log m}}\right)$. For $\sigma_1^2 = 0\left(\frac{1}{\sqrt{\log m}}\right)$, $\lim_{m\to\infty} P_2 = \alpha$, the level of significance used. Thus, corresponding to any m, however large, one can find values of σ_1^2 for which the chances are considerable (or even approaching $1 - \alpha$), that the two field trials will give opposite conclusions when the F-test is used.

13. **Asymptotic Properties of the Wald-Wolfowitz Test of Randomness.** GOTTFRIED E. NOETHER, Columbia University, New York.

Let $a_1, \cdots, a_n$ be observations on the chance variables $X_1, \cdots, X_n$. Wald and Wolfowitz (*Annals of Math. Stat.*, Vol. 14 (1943), pp. 378–388) have shown how the statistic $R_h = \Sigma_{i=1}^{n} x_i x_{i+h}$, $(x_{n+j} = x_j)$, can be used to test the null hypothesis that the X_i, $(i = 1, \cdots, n)$, are independently and identically distributed by considering the distribution of R_h in the subpopulation of all permutations of the a_i. In the present paper it is shown that when the null hypothesis is true this distribution of R_h is asymptotically normal provided $\Sigma_{i=1}^{n} (a_i - \bar{a})^r / [\Sigma_{i=1}^{n} (a_i - \bar{a})^2]^{r/2} = 0[n^{(2-r)/4}]$, $(r = 3, 4, \cdots)$, a condition which is satisfied with probability 1 if the a_i are independent observations on the same chance variable X having positive variance and a finite absolute moment of order $4 + \delta$, $(\delta > 0)$. Conditions are given for the consistency of the test based on R_h when under the alternative hypothesis observations are drawn independently from changing populations. In particular a downward trend and a regular cyclical movement are considered, both for ranks and original observations. For the special case of a regular cyclical movement of known length the asymptotic relative efficiency of the rank test with respect to the test performed on original observations is found. It is shown that when using ranks, R_h is asymptotically normal under the alternative hypothesis provided $\liminf_{n\to\infty} \operatorname{var}(n^{-5/2} R_h) > 0$. This asymptotic normality of R_h is used to compare the asymptotic power of the R_h-test with that of the Mann T-test (*Econometrica*, Vol. 13 (1945), pp. 245–259) for the case of a downward trend.

14. **On the Similar Regions of a Class of Distributions.** STEFAN PETERS, University of California, Berkeley.

The class of distributions considered is essentially the class of those distributions of n variables which, by a suitable transformation of the variables and the parameter, can be transformed into distributions defined in the whole R_n for which the parameter is a location

parameter. These regions satisfy a certain partial differential equation. The transformed distributions of the variables $y_1, \cdots y_n$ and parameter τ possess a class D_1 of similar regions with respect to τ which can be defined as the smallest additive class of regions which includes all regions defined by

$$g[(y_1 - y_n), \cdots, (y_n + y_n)] \geqq C$$

where g is a continuous function. The class D_1 does not exhaust all similar regions. There exists among the regions of class D_1 one which is most powerful for testing a given additional parameter σ. If there exists among all similar regions a most powerful region for testing σ, then that region will be the most powerful region of class D_1.

15. Some Problems in Sequential Analysis. CHARLES M. STEIN, University of California, Berkeley.

Wald's fundamental identity for cumulative sums is extended to dependent random variables. The first derivative of this at the origin is equivalent to a result of Wolfowitz (*Annals of Math. Stat.*, Vol. 18 (1947), p. 228, Th. 7.4). Higher derivatives of this at the origin can also be obtained from linear combinations of Wolfowitz's result applied to suitable products of the original random variables. These equations yield approximate *OC* and *ASN* curves for probability-ratio tests for a simple hypothesis against a single alternative concerning some of the more usual stationary Markoff chains. Bounds for the amount by which the *ASN* exceeds that of the most efficient test are also obtained. The results are applied in particular to random variables taking on only the values 0, 1 with conditional probabilities depending only on a finite number of the preceding observations. The case of linear dependence of normal random variables with fixed conditional variance is also considered.

16. Some Aspects of Links Between Prediction Problems and Problems of Statistical Estimation. ERLING SVERDRUP, University of Oslo.

A prediction is not taken as a probability statement about additional observations of the random variable already observed. It is presumed that the statistical interpretation of the sample will result in some action influencing the random variable subject to prediction. The probability distribution of this random variable is given for each of an a priori class of probability functions for the observed random variable and for each of a class of possible actions. "Utility" as a function of the random variable to be predicted and of the action is defined. It is shown that the problem of which action to take in order to maximize expected utility is identical with a problem of statistical inference with a uniquely defined weight function in the Wald sense. It is further shown that this procedure is adaptable to stochastic processes of a general type and this provides a means of connecting the theory of stochastic processes with the theory of statistical inference. Some examples are given to illustrate the general theory.

17. Some Large Sample Tests for the Median. JOHN E. WALSH, The Rand Corporation, Santa Monica, California.

Consider a large number of independent observations from continuous populations with a common median. Sone non-parametric large sample tests for the population median are presented which are based on either two or three order statistics of the sample. If all the populations are symmetrical, these tests are equal-tailed with specified significance level α. If the observations are a sample from a normal population, these tests have high power efficiencies. Some tests based on three order statistics are developed which also have signifi-

cance level α if all the populations are not symmetrical; however, in this case the resulting test is one-tailed instead of equal-tailed. Using these tests for situations where the populations are believed to be symmetrical furnishes a safety factor with respect to Type I error. Tests are presented for the special case where each population is either symmetrical or skewed in a specified direction. If the populations are not symmetrical the significance level distribution is $.4\alpha$ to one tail and $.6\alpha$ to the other, rather than $.5\alpha$ to each tail. Also some non-parametric large sample tests of whether a sample is from a symmetrical population are derived. These tests are based on three order statistics of the sample and have bounded significance levels.

18. **Continuous Sampling Plans from the Risk Point of View.** ZIVIA S. WURTELE, Stanford University, California.

The quality of a lot can be improved by a screening process whereby the defective items found during inspection are replaced by non-defective items. The type of sampling plan adopted will generally depend upon the cost of inspecting items, the number of defective items in the lot prior to inspection, and the loss due to defective items remaining in the lot after inspection. The loss if the lot is accepted after d defectives are found in a sample of n items is equal to $c(n) + h(D)$ where D is the number of defectives left in the lot and $c(n)$ is the cost of inspecting n items. An inspection procedure S is defined by a set of stopping points $\{(d, n)\}$. Let $r(p, S)$ be the expected loss if p is the probability of a defective and the procedure S is used. It is assumed that the lot is obtained from a binomial population. For any a priori distribution $F(p)$, a Bayes procedure is one which minimizes the expected risk,

$$\int_0^1 r(p,S)dF(p).$$

A systematic method of obtaining Bayes solutions exists, but the computations are formidable. Under fairly general conditions the Bayes solutions are shown to be multiple sampling plans, in which the size of the ith sample depends upon the number of defectives in the $(i - 1)$st sample. In particular, if the production is in a state of statistical control, a Bayes solution is a fixed sample size. It is also shown that for most reasonable loss functions, there exists no mini-max procedure which is uniformly better than the trivial one; namely, the Bayes procedure if $p = 1$.

NEWS AND NOTICES

Readers are invited to submit to the Secretary of the Institute news items of interest

Personal Items

Dr. Irving Burr has been promoted to a full professorship at Purdue University.

Dr. D. A. S. Fraser, who received his Ph.D. degree at Princeton University in June, has accepted a position as Instructor of Mathematics at the University of Toronto.

Dr. H. K. Hartline, formerly at the Johnsen Research Foundation of the University of Pennsylvania, has accepted an appointment as chairman of the Thomas Jenkins Department of Biophysics, Johns Hopkins University.

Dr. Leo Katz has been promoted to an associate professorship in the Mathematics Department of Michigan State College, East Lansing, Michigan.

Professor D. D. Kosambi of Tata Institute for Fundamental Research, Bombay, India served as Visiting Professor at the University of Chicago for the Winter Quarter.

Dr. H. G. Landau has resigned his position with the Ballistic Research Laboratories and is now a Research Associate with the Committee on Mathematical Biology, at the University of Chicago.

Mr. Allen L. Mayerson, formerly an Associate in the Division of Statistics and Research of the Institute of Life Insurance at New York City, has accepted a position with the National Surety Corporation of New York.

Mr. Raymond P. Peterson, who has been an Assistant in the Mathematics Department of the University of California at Los Angeles and also a graduate student there, has accepted a position with the Institute for Numerical Analysis at Los Angeles.

Professor Edwin J. G. Pitman has returned to the Mathematics Department of the University of Tasmania after spending about a year and a half in the United States. From February to June of 1948 he was at Columbia University as visiting Professor of Mathematical Statistics. The rest of the time was spent at North Carolina and Princeton.

A. Ananthapadmanabha Rau has returned to India after studying at the Statistical Laboratory in Ames, Iowa. In addition to heading the Department of Statistics and Agriculture Meterology of the Government of the State of Mysore, India, he is working on sampling design of experiments, and climatology and teaching statistics and climatology at the College of Agriculture.

Dr. Andrew Sobczyk of Watson Laboratories has been appointed to an assistant professorship at Boston University.

Assistant Professor S. L. Thompson of Alabama Polytechnic Institute has been promoted to an associate professorship.

William J. Youden is acting as Assistant Chief of the Statistical Engineering Section of the National Bureau of Standards and as special advisor to the Director on the problems of statistical and mathematical design of major experiments in physics, chemistry and engineering.

Two Doctorates in Mathematical Statistics were awarded at the University of North Carolina in June, 1949. The recipients were Uttam Chand, who has now been appointed Assistant Professor of Mathematics at Boston University, and Ralph A. Bradley, who will be Assistant Professor of Mathematics at McGill University.

The Educational Testing Service, Princeton, N. J., announces the appointment of Elbert Lee Hoffman and William Edward Kline as ETS Psychometric Fellows for 1949–50 for graduate study in psychology at Princeton University. Mr. Hoffman is a graduate of the University of Oklahoma, and Mr. Kline has received both his bachelor's and master's degree from Yale University. Bert F. Green, Jr. and Warren S. Torgerson have received reappointments as ETS Psychometric Fellows. Each Fellow carries a full program of graduate study in psychology at Princeton University, including basic work in experimental and theoretical psychology. Special training is also given in mathematical statistics and modern quantitative methods as applied to psychological problems in such fields as learning, testing and attitude measurement, as well as in the techniques of developing aptitude and achievement tests. In addition to the graduate program in psychology, each Fellow spends part-time in training and research work with the Educational Testing Service.

Preliminary Actuarial Examinations

Prize Awards

The winners of the prize awards offered by the Society of Actuaries to the nine undergraduates ranking highest on the score of Part 2 of the 1949 Preliminary Actuarial Examinations are as follows:

First Prize of $200

Moran, Joseph W.	Yale University

Additional Prizes of $100

Farmer, Thurston P., Jr.	State University of Iowa
Haakenstad, Dale L.	University of Michigan
Hauke, William V.	University of Michigan
Lordan, Joseph D.	Massachusetts Institute of Technology
Mayberry, John P.	University of Toronto
Murch, Alan D.	University of Toronto
White, William A.	Dartmouth College
Zemach, Ariel	Harvard University

The Society of Actuaries has authorized a similar set of nine prize awards for the 1950 Examinations on Part 2.

The Preliminary Actuarial Examinations consist of the following three examinations:

Part 1. *Language Aptitude Examination.*
(Reading comprehension, meaning of words and word relationships, antonyms, and verbal reasoning.)

Part 2. *General Mathematics Examination.*
(Algebra, trigonometry, coordinate geometry, differential and integral calculus.)

Part 3. *Special Mathematics Examination.*
(Finite differences, probability and statistics.)

The 1950 Preliminary Actuarial Examinations will be administered by the Educational Testing Service at centers throughout the United States and Canada on May 19, 1950. The closing date for applications is March 15, 1950.

Detailed information concerning the Examinations can be obtained from:

The Society of Actuaries
208 South LaSalle Street
Chicago 4, Illinois

New Members

The following persons have been elected to membership in the Institute

(March 1, 1949 to May 31, 1949)

Alcantara de Oliveira, Eduardo, Ph.D., (Univ. de Sao Paulo) Professor, Faculdade de Filosofia, University of Sao Paulo, *Rua Sergipe, 96-Ap. 32, Sao Paulo, Brazil.*

Ashby, Wallace L., A.B. (George Washington Univ.) Agricultural Statistician, *3746 Jocelyn Street, Washington 15, D. C.*

Bailey, Edward W., B.Ch. (Ohio State Univ.) Quality Control Supervisor, Carbide and Carbon Chemicals Corporation, Y-12 Plant, *101 Moylan Lane, Oak Ridge, Tennessee.*

Berger, Agnes P., Ph.D. (Budapest) *10 Park Avenue, New York, New York.*

Brown, Walter C., B.S. (Colorado A&M College) Graduate Assistant, Department of Mathematics, University of Oklahoma, *1130 Trout, Norman, Oklahoma.*

Calvin, Lyle D., B.S. (Univ. of Chicago) Research Graduate Assistant, Institute of Statistics, North Carolina State College, Raleigh, North Carolina.

Carlyle, Charles G., B.S. (Univ. of Illinois) Graduate student at University of Illinois, *C-32 Stadium Terrace, Champaign, Illinois.*

Chen, Yu-nien, M.A. (Harvard) Graduate student, Harvard University, *147-60, Apt. D, Charter Road, Jamaica 2, New York.*

Clark, Fred J., Jr., B.S. (Colorado A&M College) Graduate Assistant at University of Illinois, Department of Mathematics, *61 A Court G, Stadium Terrace, Champaign, Illinois.*

Cohen, Samuel E., M.A. (Univ. of Pennsylvania) Statistician, U. S. Bureau of Labor Statistics, *49 Galveston St., S.W., Washington 20, D. C.*

Cole, Randal H., Ph.D. (Univ. of Wisconsin) Associate Professor, University of Western Ontario, London, Canada.

Comrey, Andrew L., Ph.D. (Univ. of Southern Calif.) Assistant Professor of Psychology, University of Illinois, Urbana, Illinois.

Cook, Ellsworth B., B.S. (Springfield College) Head of Visual Screening Devices Research and Statistics Facility, U. S. Naval Medical Research Laboratory, *Box 45, Submarine Base, New London, Connecticut.*

Cox, David R., Ph.D. (Leeds, England) Statistician, Wool Industries Research Association, *2 Sunset Avenue, Leeds 6, Yorks, England.*

Denbow, Carl H., Ph.D. (Univ. of Chicago) Associate Professor of Mathematics, U. S. Naval Postgraduate School, Annapolis, Maryland.

Dillon, Gregory M., A.B. (Long Island Univ.) Statistician, Pension Statistics Section,

Treasury Department, E. I. DuPont de Nemours & Co., *1331 Cedar Street, Wilmington, Delaware.*

Duarte, Geraldo Garcia, Licenciado em Matematica (Faculdade de Filosofia de S. Bento) Assistente da Faculdate de Higiene e Saude Publica, Caixa Postal 99B, Sao Paulo, Brazil.

Dudman, John A., B.A. (Reed College) Graduate student, Columbia University, *56 West 70th St., New York 23, New York.*

Edelson, Howard, B.A. (Ohio State Univ., Columbus, Ohio) Graduate student and Graduate Assistant, Ohio State University, *794 S. 18th St., Columbus 6, Ohio.*

Feron, R., Licencie es Sciences, (Univ. of Paris) Attache de Recherce, *13 rue des Feuillantines, Paris V, France.*

Franck, Edward Michel, Inpineur A.I.A., Professor of the Royal Military School, *104 Rue Pere Devroye, Woluwe St. Pierre, Belgium.*

Garritsen, Florence M., B.A. (Univ. of Michigan) Research Assistant, General Motors Corp., *5151 Lillibridge Ave., Detroit 13, Michigan.*

Gelsomini, Thea, Ph.D. (Univ. of Bocconi, Milano) Assistician of Statistics at Department of Statistics, University of Bocconi, *Via A. Stoppi, N. 10, Milano, Italy.*

Goudswaazd, G, Ph.D. (Univ. of Leiden) Director, Permanent Office, International Statistical Institute and Lecturer of Statistics, Rotterdam School of Economics and Free University of Amsterdam, 2 Oostduinlaan, The Hague, Netherlands.

Gucker, Frank Fulton, A.B. (Harvard Univ.) Statistical Engineer, Remington Arms Co., Inc., *3175 Main Street, Bridgeport 6, Connecticut.*

Haberman, Sol, B.A. (Brooklyn College) Assistant Visiting Professor of Sociology, University of Puerto Rico, *187 Avenida los Flamboyanes, Rio Piedras, Puerto Rico.*

Heimbach, Ernest E., M.B.A. (New York Univ.) Professor of Economics, Bergen College, Teaneck, New Jersey, *55 West 11th Street, New York 11, N. Y.*

Ishii, Shigeru, B.A. (Univ. of Ill.) Student at University of Illinois, *320-1 Peabody Drive, Parade Ground Units, Champaign, Illinois.*

Jackson, James Edward, M.A. (Univ. of N.C.) Statistician, Color Control Dept., Eastman Kodak Company, *200 Pershing Drive, Rochester, New York.*

Jaspen, Nathan, Ph.D. (Pennsylvania State College) Research Assistant, Department of Psychology, Pennsylvania State College, State College, Pennsylvania.

Jonhagen, Sven, Fil Lic. (Univ. of Stockholm) Chief Actuary and Assistant Teacher in Statistics at the University of Stockholm, *Tegnergatan 36, Stockholm, Sweden.*

Kiefer, Jack C., M.S. (Mass. Inst. of Tech.) Student, Department of Mathematical Statistics, Columbia University, *3826 Middleton Avenue, Cincinnati 20, Ohio.*

Kraft, Charles Hall, B.A. (Mich. State College) Instructor, Mathematics Department, Michigan State College, *707D Chestnut Road, East Lansing, Mich.*

Mayne, John W., M.Sc. (Brown Univ.) Graduate student in Mathematical Statistics, 330 Furnald Hall, Columbia University, New York 27, New York.

McCabe, William J., M.A. (George Washington Univ.) Chief Statistician, Transportation Corps, Department of the Army, *1725 South Oakland Street, Arlington, Virginia.*

Medin, Knut H., M.A. (Univ. of Uppsala) Assistant, Statistical Institute, University of Uppsala, Odinslund 2, Uppsala, Sweden.

Mewborn, A. Boyd, Ph.D. (Calif. Inst. of Tech.) Associate Professor of Mathematics and Mechanics, *P.O. Box 1748, Monterey, California.*

Minton, Paul D., M.S. (Southern Methodist Univ.) Graduate student, University of North Carolina, P.O. Box 634, Chapel Hill, North Carolina.

Morris, Doris N., M.A. (Columbia Teachers College) Economics Assistant, Western Electric Co., *101 West 72nd St., New York 23, New York.*

Morris, Robert H., B.A. (Swarthmore) Development Engineer, Color Control Department, Eastman Kodak Co., Rochester, New York.

Rajalakshman, D. V., M.Sc. (Madras Univ.) Head, Department of Statistics, University of Madras, Madras 5, S. India.

Rudy, Norman, M.B.A. (Univ. of Chicago) Scientist, Ordnance Research Project, University of Chicago and Instructor in Economics, Roosevelt College, *7105 S. Crandon Avenue, Chicago 49, Illinois.*

Sakk, Kaarel, Fil. kand. (Univ. of Stockholm) Officer at Research Bureau of the State Foodstuffs Commission, Ostermalmsgatan 67 o.g. III, Stockholm, Sweden.

Singh, Jagjit, B.A. (Punjab Univ.) Superintendent Transportation, E. I. Railway, Dinapore, *c/o B.S. Bugga Esqr., Post Office Box 441, Calcutta, India.*

Starr, Henry H., Ph.D. (Univ. of Vienna) Research Manager, Converted Rice, Inc., P.O. Box 1752, Houston 1, Texas.

Sverdrup, Erling, Actuarian (Univ. of Oslo) Lecturer in Mathematical Statistics, Institute of Mathematics, University of Oslo, Oslo, Norway.

Talacko, Joseph Y., Ph.D. (Charles Univ., Prague) Assistant Professor of Mathematics, Marquette University, *2503 So. 10th Street, Milwaukee 7, Wisconsin.*

Templeton, James G. C., A.M. (Princeton Univ.) Graduate student at Princeton University, Fine Hall, Princeton University, Princeton, New Jersey.

Vaughan, Elizabeth, B.S. (Univ. of Washington) Statistician, U.S. Fish and Wildlife Service, 2725 Montlake Boulevard, Seattle 2, Washington.

Wilkinson, Bryan, M.A. (Univ. of Nebraska) Personnel Research Specialist, Prudential Insurance Co., Western Home Office, *4130-B Muirfield Road, Los Angeles 43, California.*

Yevick, Mariam A. L., Ph.D. (Mass. Inst. of Tech.) Staff, Division of Statistical Engineering, National Broadcasting System, *921 Hudson St., Hoboken, New Jersey.*

REPORT ON THE BERKELEY MEETING OF THE INSTITUTE

The thirty-ninth meeting and fifth regional West Coast meeting of the Institute of Mathematical Statistics was held on the Berkeley campus of the University of California, from Thursday June 16 through Saturday June 18, 1949. The session on June 17 was held jointly with the Biometrics Section of the American Statistical Association and the Biometric Society (Western N. A. Region). Sixty-six persons registered, including the following fifty members of the Institute:

Jane F. Andrian, G. A. Baker, Z. Wm. Birnbaum, Colin R. Blyth, Albert H. Bowker, Paul T. Bruyere, Chin Long Chiang, Edwin L. Crow, John H. Curtiss, R. C. Davis, Carl H. Denbow, W. J. Dixon, Mary Elveback, Mark Eudey, Edward A. Fay, Evelyn Fix, William R. Gaffey, H. H. Germond, M. A. Girshick, Jack Gysbers, Max Halperin, J. L. Hodges, Jr., John M. Howell, Harry M. Hughes, Cuthbert Hurd, Terry A. Jeeves, Mark Kac, H. S. Konijn, George M. Kuznets, Erich L. Lehmann, Richard F. Link, Michel Loève, Frank Massey, Lincoln E. Moses, Edith Mourier, Stanley W. Nash, J. Neyman, Edward Paulson, Stefan Peters, Raymond P. Peterson, Robert I. Piper, Gladys Rappaport, Mina Rees, David Rubinstein, Elizabeth L. Scott, Esther Seiden, Charles M. Stein, John E. Walsh, John Wishart, Zivia S. Wurtele.

Those attending were welcomed at the Thursday morning session by Edward W. Strong, Associate Dean of the College of Letters and Science, University of California. Professor Z. William Birnbaum of the University of Washington presided.

The program was as follows:

1. *Recent advances in the theory of the Wishart distribution.* (Invited paper.) John Wishart, Cambridge University.
2. *Bayes, minimax, and other approaches to the multiple classification problem.* (Invited paper.) M. A. Girshick, Stanford University.
3. *Some problems in sequential analysis.* Charles M. Stein, University of California, Berkeley.

Professor Jerzy Neyman of the University of California, Berkeley, presided at the Thursday afternoon session. Midway in the program there was an intermission for a tea given by the Statistical Laboratory, University of California. The program was as follows:

1. *Completeness in the sequential case.* E. L. Lehmann and C. M. Stein, University of California, Berkeley.
2. *Some large sample tests for the median.* John E. Walsh, The Rand Corporation.
3. *Continuous sampling plans from the risk point of view.* Zivia S. Wurtele, Stanford University.
4. *Some problems in point estimation.* J. L. Hodges, Jr. and E. L. Lehmann, University of California, Berkeley.
5. *Minimum variance in non-regular estimation.* R. C. Davis, U. S. Naval Ordnance Test Station, Inyokern.
6. *Some aspects of links between prediction problems and problems of statistical estimation.* Erling Sverdrup, University of Oslo.

7. *Extension of a theorem of Blackwell.* (By title). Edward W. Barankin, University of California, Berkeley.
8. *Some two-sample tests.* (By title). Douglas G. Chapman, University of California, Berkeley.
9. *On the existence of consistent tests.* (By title). Agnes Berger, Columbia University.

Professor F. W. Weymouth of Stanford University presided at the Friday morning session on biometrics. The program was as follows:

1. *Statistical problems arising from research in tuberculosis.* Martha and Paul T. Bruyere, U. S. Public Health Service.
2. *Correlation of variability with growth rate in fish and mollusks.* F. W. Weymouth, Stanford University.
3. *Some problems arising in plant selection and the use of analysis of variance.* Stanley W. Nash, University of California, Berkeley.
4. *Studies of resistance of strawberry varieties and selections to verticillum wilt.* R. E. Baker and G. A. Baker, University of California, Davis.
5. *A uniformity trial on unirrigated barley of ten years duration with implications for field trial designs.* F. J. Veihmeyer, M. R. Huberty, and G. A. Baker, University of California, Davis and Los Angeles.

On Friday afternoon those attending the meeting were entertained at a picnic luncheon at Stanford University, given by the Department of Statistics, Stanford University.

Professor C. B. Morrey, Jr., of the University of California, Berkeley, presided at the Saturday morning session. The program consisted of the following invited papers:

1. *Methods for getting limiting distributions.* Mark Kac, Cornell University.
2. *Almost certain convergence.* Michel Loève, University of California, Berkeley.

At 11 o'clock Saturday morning a business session was held, under the chairmanship of Professor Jerzy Neyman of the University of California, Berkeley, for the purpose of discussing future West Coast meetings. Plans for reviving the Statistical Research Memoirs were also discussed.

On Saturday afternoon a final session for contributed papers was held under the chairmanship of Professor Albert H. Bowker of Stanford University. The program was as follows:

1. *Effect of linear truncation in a multinormal population.* Z. William Birnbaum, University of Washington.
2. *Estimation in truncated samples.* Max Halperin, The Rand Corporation.
3. *On the similar regions of a class of distributions.* Stefan Peters, University of California, Berkeley.
4. *Auxiliary random variables.* Mark W. Eudey, California Municipal Statistics.
5. *The ratio of ranges.* Richard F. Link, University of Oregon.
6. *Statistical problems in the theory of Geiger counters.* Colin R. Blyth, University of California, Berkeley.
7. *Asymptotic properties of the Wald-Wolfowitz test of randomness.* (By title). Gottfried E. Noether, Columbia University.

J. L. Hodges, Jr.
Assistant Secretary